TESTING OF HOMOGENEITY IN DISTRIBUTIONS WITH ORDERED CATEGORIES

by

LEE Chi-ming

A Thesis Submitted to the

Graduate School

of

The Chinese University of Hong Kong

(Division of Statistics)

In Partial Fulfillment

of the Requirements for the Degree of

Master of Philosophy

(M. Phil.)

June, 1995

THE CHINESE UNIVERSITY OF HONG KONG

GRADUATE SCHOOL

The undersigned certify that we have read a thesis, entitled "Testing of Homogeneity in Distributions with Ordered Categories" submitted to the Graduate School by Lee Chi Ming (李 智 明) in partial fulfillment of the requirement for the degree of Master of Philosophy in Statistics. We recommend that it be accepted.

W. Y. Poon

Dr. W.Y. Poon,

Supervisor.

Lee Sik Yum

Dr. S.Y. Lee

Kim-Hung Li

Dr. K.H. Li

Prof. P.M. Bentler,

External Examiner.

DECLARATION

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

ACKNOWLEDGEMENT

I would like to express my deepest and earnest gratitude to my supervisor, Dr. W. Y. Poon. Her encouragement and kind supervision have helped me a lot during the preparation of this thesis.

LEE Chi-ming

Department of Statistics

The Chinese University of Hong Kong

June, 1995.

ABSTRACT

In categorical data analysis, the chi-square test is one of the most common method to deal with the test of homogeneity. However, it may be inappropriate in the case of ordinal categorical data. Hence, a new approach is suggested to analyze this kind of data.

For simplicity, suppose we have two sets of observable ordinal categorical data, it is assumed that each comes from the same univariate distribution. Then two continuous variables are obtained. They are supposed to share the same thresholds and the same number of categories. One of the variables is considered as the reference variable and the value of the parameter determining the underlying univariate distribution is given. By maximum likelihood method, the thresholds can be estimated via Newton-Raphson algorithm. By considering the same univariate distribution imposed on the reference variable, the value of the parameter in the distribution of the other variable can also be estimated by maximum likelihood method. Then it is possible to compare two distributions in a relative sense by setting up useful hypotheses on the parameters involved.

In Chapter 3, the goodness-of-fit test is set up to determine which one of the univariate distributions should be chosen and imposed. A real data is considered to illustrate the proposed method in Chapter 4.

Contents

1 Introduction **1**

2 Three Underlying Distributions **6**

2.1 Exponential distribution 8

2.1.1 Estimation of Thresholds 8

2.1.2 Estimation of Parameter 11

2.1.3 Simulation Study 13

2.2 Normal distribution 24

2.2.1 Estimation of Thresholds 24

2.2.2 Estimation of Parameter 26

2.2.3 Simulation Study 29

2.3 Weibull distribution 42

2.3.1 Estimation of Thresholds 42

2.3.2 Estimation of Parameter 44

2.3.3 Simulation Study 46

3 Goodness-of-fit Test **60**

3.1 Test for the Exponential distribution 63

3.2 Test for the Normal distribution 66

3.3 Test for the Weibull distribution 68

3.4 Implication 70

3.5 An Artificial Example 72

3.5.1 Case 1 ($s = 3$) 72

3.5.2 Case 2 ($s = 4$) 80

4 Real Data Illustration **87**

4.1 Test for the Exponential distribution 90

4.2 Test for the Normal distribution 91

4.3 Test for the Weibull distribution 92

4.4 Inferences from the Exponential distribution 94

4.5 Inferences from the Normal distribution 95

4.6 Inferences from the Weibull distribution 98

5 Conclusion **101**

Bibliography **103**

List of Tables

2.1 Cross-Classification of Respondents in 1972-1975. 6

2.2 Simulation Study on Exponential Distribution. Case (A). 19

2.3 Simulation Study on Exponential Distribution. Case (B). 20

2.4 Simulation Study on Exponential Distribution. Case (C). 21

2.5 Simulation Study on Normal Distribution. Case (A). 31

2.6 Simulation Study on Normal Distribution. Case (B). 34

2.7 Simulation Study on Normal Distribution. Case (C). 37

2.8 Simulation Study on Weibull Distribution. Case (A). 49

2.9 Simulation Study on Weibull Distribution. Case (B). 52

2.10 Simulation Study on Weibull Distribution. Case (C). 55

3.1 Artificial 3×3 Contingency Table. 73

3.2 Artificial 3×4 Contingency Table. 81

4.1 Modified Data Set with Ordered Responses. 88

4.2 Hypothesis Testing on parameter λ. 94

4.3 Parameter Estimates with Estimated Standard Errors (S.E.). . . . 96

4.4 Hypothesis Testing on parameters, μ and σ. 97

4.5 Parameter Estimates with Estimated Standard Errors (S.E.). . . . 98

4.6 Hypothesis Testing on parameters, γ and β. 99

Chapter 1

Introduction

In social sciences, ordinal scales are commonly occurred, in particular for measuring attitudes and opinions on various issues and status of various types. As an illustration of ordinal variables and the levels of their corresponding scales, political philosophy may be classified as "liberal", "moderate", or "conservative"; social class may be measured as "upper", "middle", or "lower"; opinion on abortion may have responses as "should be available on demand", "should only be allowed in particular circumstances", or "should never be allowed" and so on. Though ordinal scales are common in social sciences, they are no means restricted to those areas. They occur frequently in behavioral sciences, public health, education and marketing.

In the General Social Surveys [7], many of the same questions are asked

from year to year. To examine whether there is a trend over the years of survey, the test of homogeneity of several independent samples may be questioned in advance. As an example, extracting from the General Social Surveys, subjects have been asked whether courts were sufficiently harsh with criminals from year 1972 to 1975. Responses are classified as "Not harshly enough", "About right" and "Too harshly". Given the responses from the subjects, overall changes in attitudes toward the courts can be investigated.

Basically, the Pearson chi-squared statistic is commonly used to test homogeneity. However, the Pearson chi-squared test may have some restrictions or drawbacks as the sample size should be large enough in order for the chi-squared distribution to give a good approximation for the exact sampling distribution of the chi-squared statistic. Secondly, in analyzing categorical data with ordinal variables, the Pearson chi-squared test of homogeneity is not appropriate since the chi-squared statistics for testing homogeneity are invariant to permutations of rows and permutations of columns. So, it may ignore some of the available information given by the ordinal categories. On the other hand, suppose the case of homogeneity holds, we come to a conclusion that attitudes toward the courts are not changing over time. However, if the null hypothesis of homogeneity is rejected, we can not, as if in the continuous case, attribute the source of heterogeneity to location difference or variation difference.

Ordinal categorical variables do not have origins or units of measurements. It is generally nonsense to consider mean and variance of an ordinal variable. However, for ordinal categorical variables with comparable categories, it is intuitively possible to consider their means and variances in a relative sense. For more discussion on this context, the following will give detailed explanation.

Let Z be an observable ordinal categorical variable. It is assumed that it is related to an underlying continuous variable X by :

$$Z = k \qquad \text{if} \quad \alpha_k \leq X < \alpha_{k+1} \qquad \text{for} \quad k = 1, \ldots, s, \tag{1.1}$$

where s is the number of categories of Z and $\alpha_1, \alpha_2, \ldots, \alpha_{s+1}$ are the thresholds of Z with known α_1 and α_{s+1}. Suppose we consider two continuous variables X_1 and X_2, having the same number of categories and sharing the same thresholds. If we take X_1 as the reference variable and impose a value on the parameter in the underlying distribution of X_1. Since X_1 and X_2 possess comparable categories, it is possible to study the distribution of X_2 relatively by estimating the value of the parameter in the same distribution considered.

More precisely, if X_1 is considered as the reference variable and a univariate distribution is imposed to X_1. The value of the parameter which determines the univariate distribution is given. By maximum likelihood estimation, the threshold estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha} = \{\alpha_2, \alpha_3, \ldots, \alpha_s\}$ can be obtained. Similarly, by considering the same univariate distribution imposed on X_1, the value of the parameter in the

distribution of X_2 can be estimated given the thresholds fixed at $\hat{\boldsymbol{\alpha}}$. Therefore, it is possible to compare two distributions in a relative sense by setting up useful hypotheses on the parameters involved.

In Chapter 2 of this thesis, three distributions are mainly considered, that is, exponential, normal and Weibull distributions. They are applied in a very wide variety of statistical procedures. The exponential distribution is frequently used in the field of life-testing. The lifetime can often be usefully represented by an exponential random variable with a relatively simple associated theory. If the representation is not adequate, a modification of the exponential distribution (very often a Weibull distribution) is used. The Weibull distribution may provide the extra flexibility to make a model sufficiently accurate for use in an analysis. The last but not the least, the normal distribution holds a central position in statistics. It has the familiar bell shape, whose symmetry makes it an appealing choice for many population models. The details are discussed in Balakrishnan, Johnson and Kotz [3]. Based on the above underlying distributions, three sections are involved. In the first part of each section, given the observed data and a value on the parameter which determines the univariate distribution, the thresholds are estimated by maximum likelihood estimation via Newton-Raphson algorithm. In the second part, given the thresholds, the estimation of parameter is discussed. In addition, a simulation study on the test of homogeneity is presented according

to the different choice of parameter value of the distribution, the choice of the true thresholds, and the choice of the sample size in each section.

In Chapter 3, goodness-of-fit test is considered to determine which one of the three distributions should be chosen and imposed. Although only three distributions are mainly considered in this thesis, it doesn't mean that all cases are restricted to them. Many other distributions may be possible to be considered. In § 3.4, the goodness-of-fit test is discussed in a more general way. Chapter 4 gives a real data illustration which is extracted from the General Social Surveys to show how the proposed method can be applied to give some insights in analyzing ordinal categorical data. Lastly, Chapter 5 comes to a conclusion with a precise summary of whole thesis.

Chapter 2

Three Underlying Distributions

Consider the following data extracted from the General Social Surveys :

Table 2.1: Cross-Classification of Respondents in 1972-1975.

	Year of survey				
Response	1972	1973	1974	1975	Total
Too harshly	105	68	42	61	276
About right	265	196	72	144	677
Not harshly enough	1066	1092	580	1174	3912
Don't know	173	138	51	104	466
No answer	4	10	8	7	29
Total	1613	1504	753	1490	5360

Table 2.1 shows that from year 1972 to 1975, subjects were asked whether courts were sufficiently harsh with criminals. The main theme in this thesis is concentrated on testing homogeneity among the years of survey. Since the ordinal scales are considered, the categories of the response variable are selected and only three categories are remained as "Not harshly enough", "About right" and "Too harshly". Therefore, the thresholds considered are α_2, α_3 with known α_1 and α_4. From the four years, we take year 1972 as the reference year without loss of generality. If a univariate distribution is imposed to the reference year and the value of the parameter which determines the distribution is given. By maximum likelihood estimation, the threshold estimates of α_2, α_3 can be obtained. Next, by considering the same univariate distribution, given the thresholds fixed at the threshold estimates found before, the value of the parameter in the distribution of year 1973, year 1974 or year 1975 can also be found. In this chapter, we consider three kinds of distributions, that is, exponential distribution, normal distribution and Weibull distribution. Based on the three underlying distributions, three sections are involved. In each section, the estimation of thresholds, the estimation of parameter and the simulation study are provided.

2.1 Exponential distribution

Let Z be an observable random variable, relating to an underlying continuous variable X by :

$$Z = k \qquad \text{if} \quad \alpha_k \leq X < \alpha_{k+1} \quad \text{for} \quad k = 1, \ldots, s, \tag{2.1}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_{s+1}$ are the thresholds of Z with $\alpha_1 = 0$ and $\alpha_{s+1} = +\infty$. It is assumed that the variable X has an exponential distribution of the density function :

$$f(x) = f(x; \lambda) = \frac{1}{\lambda} \exp\left\{-\frac{x}{\lambda}\right\}, \qquad x \geq 0, \quad \lambda > 0. \tag{2.2}$$

The only one parameter λ is known as the scale parameter, since most of its influence is on the spread of the distribution. The mean and variance of the exponential distribution are λ and λ^2 respectively.

2.1.1 Estimation of Thresholds

In this part, given the observed data and the parameter λ, the objective is to estimate the thresholds $\alpha_2, \ldots, \alpha_s$ by maximum likelihood method. Suppose a random sample of size n is observed. The likelihood function is given by :

$$L(\boldsymbol{\alpha}) \;=\; c \prod_{i=1}^{s} p_i(\boldsymbol{\alpha})^{n_i}, \tag{2.3}$$

where c is a constant independent of the parameter vector $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_s)'$, n_i is the observed frequency in the i-th cell for $i = 1, 2, \ldots, s$ with $\sum_{i=1}^{s} n_i = n$ and p_i is the probability of an observation falling in the i-th cell.

Then

$$p_1 = \Pr(Z = 1) = \int_{\alpha_1}^{\alpha_2} f(x; \lambda)\, dx,$$

$$\vdots$$

$$p_s = \Pr(Z = s) = \int_{\alpha_s}^{\alpha_{s+1}} f(x; \lambda)\, dx. \tag{2.4}$$

The maximum likelihood estimate (MLE) $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ is the vector which maximizes the likelihood function $L(\boldsymbol{\alpha})$, or equivalently, minimizes the negative log-likelihood function :

$$F(\boldsymbol{\alpha}) = -\log\, L(\boldsymbol{\alpha}) \;\propto\; -\sum_{i=1}^{s} n_i\, \log\left\{\int_{\alpha_i}^{\alpha_{i+1}} f(x; \lambda)\, dx\right\}. \tag{2.5}$$

To obtain the maximum likelihood estimate $\hat{\boldsymbol{\alpha}}$, it is required to solve the equations :

$$\frac{\partial F(\boldsymbol{\alpha})}{\partial \alpha_j} = 0, \qquad j = 2, 3, \ldots, s.$$

In general, the minimum cannot be found in closed form. The classical Newton-Raphson algorithm is used to minimize the negative log-likelihood function. The basic step of the Newton-Raphson algorithm is given by :

$$\triangle\boldsymbol{\alpha} = -\gamma H(\boldsymbol{\alpha})^{-1} \dot{F}(\boldsymbol{\alpha}), \tag{2.6}$$

where γ is a step-size parameter which may be taken as the first value in the sequence $1, \frac{1}{2}, \frac{1}{4}, \ldots$ that reduces F.

$$\dot{F}(\boldsymbol{\alpha}) = \frac{\partial F(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \tag{2.7}$$

is the gradient vector and

$$H(\boldsymbol{\alpha}) = \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \tag{2.8}$$

is the Hessian matrix. Refer to (2.5), the negative log-likelihood function $F(\boldsymbol{\alpha})$ is derived as :

$$F(\boldsymbol{\alpha}) \propto -\sum_{i=1}^{s} n_i \log\left\{\int_{\alpha_i}^{\alpha_{i+1}} f(x)\, dx\right\},$$

where

$$f(x) = f(x;\lambda) = \frac{1}{\lambda} \exp\left\{-\frac{x}{\lambda}\right\}.$$

The first and second derivatives are computed as follows :

$$\frac{\partial F(\boldsymbol{\alpha})}{\partial \alpha_j} = \frac{e^{-\alpha_j/\lambda}}{\lambda}\left\{\frac{n_j}{\int_{\alpha_j}^{\alpha_{j+1}} f(x)dx} - \frac{n_{j-1}}{\int_{\alpha_{j-1}}^{\alpha_j} f(x)dx}\right\} \qquad \text{for } j = 2, \ldots, s, \tag{2.9}$$

$$\begin{aligned}\frac{\partial^2 F(\boldsymbol{\alpha})}{\partial {\alpha_j}^2} &= \frac{n_{j-1} e^{-\alpha_j/\lambda}}{\lambda^2 \int_{\alpha_{j-1}}^{\alpha_j} f(x)dx}\left\{1 + \frac{e^{-\alpha_j/\lambda}}{\int_{\alpha_{j-1}}^{\alpha_j} f(x)dx}\right\} \\ &\quad - \frac{n_j e^{-\alpha_j/\lambda}}{\lambda^2 \int_{\alpha_j}^{\alpha_{j+1}} f(x)dx}\left\{1 - \frac{e^{-\alpha_j/\lambda}}{\int_{\alpha_j}^{\alpha_{j+1}} f(x)dx}\right\} \qquad \text{for } j = 2, \ldots, s.\end{aligned} \tag{2.10}$$

The above second derivatives are the diagonal entries of the Hessian matrix. Also, some of the non-zero off-diagonal entries are computed as :

$$\frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_{j-1} \partial \alpha_j} = - \frac{n_{j-1} e^{-\frac{\alpha_j}{\lambda}} e^{-\frac{\alpha_{j-1}}{\lambda}}}{\lambda^2 (\int_{\alpha_{j-1}}^{\alpha_j} f(x)dx)^2} \qquad \text{for } j = 3, \ldots, s,$$

$$\frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_{j+1} \partial \alpha_j} = - \frac{n_j e^{-\frac{\alpha_j}{\lambda}} e^{-\frac{\alpha_{j+1}}{\lambda}}}{\lambda^2 (\int_{\alpha_j}^{\alpha_{j+1}} f(x)dx)^2} \qquad \text{for } j = 2, \ldots, s-1. \tag{2.11}$$

Apart from the above entries of the Hessian matrix, all the other entries are equal to zero. More precisely, the Hessian matrix is :

$$H(\boldsymbol{\alpha}) = \begin{pmatrix} \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial {\alpha_2}^2} & \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_2 \partial \alpha_3} & 0 & 0 & 0 & \cdots & 0 \\ \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_3 \partial \alpha_2} & \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial {\alpha_3}^2} & \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_3 \partial \alpha_4} & 0 & 0 & \cdots & 0 \\ 0 & \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_4 \partial \alpha_3} & \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial {\alpha_4}^2} & \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_4 \partial \alpha_5} & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \ddots & \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial {\alpha_{s-1}}^2} & \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_{s-1} \partial \alpha_s} \\ 0 & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_s \partial \alpha_{s-1}} & \frac{\partial^2 F(\boldsymbol{\alpha})}{\partial {\alpha_s}^2} \end{pmatrix} \tag{2.12}$$

2.1.2 Estimation of Parameter

In the second part, the thresholds are assumed to be the given constants, the parameter λ in the exponential distribution is estimated by maximum likelihood method. The maximum likelihood estimate $\hat{\lambda}$ of λ is one which maximizes the

likelihood function $L(\lambda)$, the same as one stated in previous part, or equivalently, minimizes the negative log-likelihood function :

$$G(\lambda) = -\log\ L(\lambda)\ \propto\ -\sum_{i=1}^{s} n_i\ \log\left\{\int_{\alpha_i}^{\alpha_{i+1}}\ f(x;\lambda)\ dx\right\}. \tag{2.13}$$

The classical Newton-Raphson algorithm is also used to find the minimum. Therefore, the first and second derivatives are required. Derivatives of $G(\lambda)$ are found as follows.

The first derivative is derived as :

$$\begin{aligned}
\frac{\partial G(\lambda)}{\partial \lambda} &= \frac{\partial}{\partial \lambda}\left\{\sum_{i=1}^{s} -\ n_i\ \log \int_{\alpha_i}^{\alpha_{i+1}} f(x;\lambda)dx\right\} \\
&= \sum_{i=1}^{s} \frac{-\ n_i}{\int_{\alpha_i}^{\alpha_{i+1}} f(x;\lambda)dx}\ \frac{\partial}{\partial \lambda}\int_{\alpha_i}^{\alpha_{i+1}} \frac{1}{\lambda} e^{-x/\lambda} dx \\
&= \sum_{i=1}^{s} n_i \left\{\frac{1}{\lambda} - \frac{\int_{\alpha_i}^{\alpha_{i+1}} x e^{-x/\lambda} dx}{\lambda^2 \int_{\alpha_i}^{\alpha_{i+1}} e^{-x/\lambda} dx}\right\}.
\end{aligned} \tag{2.14}$$

The second derivative is derived as :

$$\frac{\partial^2 G(\lambda)}{\partial \lambda^2} = \sum_{i=1}^{s} \frac{n_i}{\lambda^2}\left\{-1 - \frac{\int_{\alpha_i}^{\alpha_{i+1}} x^2 e^{-x/\lambda} dx}{\lambda^2 \int_{\alpha_i}^{\alpha_{i+1}} e^{-x/\lambda} dx} + \frac{\left(\int_{\alpha_i}^{\alpha_{i+1}} x e^{-x/\lambda} dx\right)^2}{\lambda^2 \left(\int_{\alpha_i}^{\alpha_{i+1}} e^{-x/\lambda} dx\right)^2} + \frac{2\int_{\alpha_i}^{\alpha_{i+1}} x e^{-x/\lambda} dx}{\lambda \int_{\alpha_i}^{\alpha_{i+1}} e^{-x/\lambda} dx}\right\}. \tag{2.15}$$

It is well kwown that under mild regularity conditions, the maximum likelihood estimate $\hat{\lambda}$ of λ is consistent. According to the asymptotic theory (See, e.g., Rao [11]), it can be shown that,

$$\sqrt{n}\ (\hat{\lambda} - \lambda)\ \xrightarrow{l}\ N[0,\ \mathrm{I}^{-1}(\lambda)], \tag{2.16}$$

where l denotes convergence in distribution. $\mathrm{I}(\lambda)$ is the information matrix defined as :

$$\mathrm{I}(\lambda) = \mathrm{E}\left\{(\frac{\partial G(\lambda)}{\partial \lambda})\,(\frac{\partial G(\lambda)}{\partial \lambda})'\right\} = \mathrm{E}\left\{\frac{\partial^2 G(\lambda)}{\partial \lambda^2}\right\}. \tag{2.17}$$

Since the Hessian matrix converges in probability to $\mathrm{I}(\lambda)$, the Hessian matrix can be used to approximate the information matrix. In this case, the estimate of the variance of $\hat{\lambda}$ is given by $(H(\hat{\lambda}))^{-1}$ where $H(\lambda) = \partial^2 G(\lambda)/\partial \lambda^2$. Therefore, by the Newton-Raphson algorithm, the asymptotic covariance matrix of $\hat{\lambda}$ can be estimated, and hence the estimated standard errors will also be obtained.

2.1.3 Simulation Study

In order to investigate the performance of the above proposed algorithm, a simulation study is set up in this section. By varying the value of the parameter, the true value of the thresholds and the sample size, the performance of the estimates are discussed. For each set of values, the number of replications is 100 and the root mean square of the gradients is used as the convergence criterion. The iteration stops when the root mean square is smaller than a pre-assigned value, say δ=0.00001.

This study is based on simulating data from two exponential distributions. One is the reference group and the other is the one under the test of homogeneity.

Based on the true thresholds, two grouped frequencies are obtained. In the reference group, the thresholds can be estimated by maximum likelihood method. The detailed procedure is given in § 2.1.1. Then the mean of the threshold estimates is given by :

$$\bar{\alpha}_i = \frac{1}{100}\sum_{k=1}^{100} \hat{\alpha}_i(k) \quad \text{for } i = 2, 3, \ldots, s, \tag{2.18}$$

where $\hat{\alpha}_i(k)$ is the estiamte of the $(i-1)$-th component of the threshold parameter $\boldsymbol{\alpha} = (\alpha_2, \alpha_3, \ldots, \alpha_s)'$ in the k-th replication. Moreover, the root-mean-square error of the threshold estimates (RMS_i) is obtained as :

$$\text{RMS}_i = \left\{\frac{1}{100}\sum_{k=1}^{100} (\hat{\alpha}_i(k) - \alpha_i)^2\right\}^{\frac{1}{2}} \quad \text{for } i = 2, 3, \ldots, s, \tag{2.19}$$

where α_i is the true value of the $(i-1)$-th component of the threshold parameter $\boldsymbol{\alpha}$.

Based on the thresholds estimated in the reference group, the parameter λ in the exponential distribution can be estimated by maximum likelihood method in second group's data. The procedure is presented in § 2.1.2. The mean of the parameter estimate is given by :

$$\bar{\lambda} = \frac{1}{100}\sum_{k=1}^{100} \hat{\lambda}(k), \tag{2.20}$$

where $\hat{\lambda}(k)$ is the estimate of the parameter in the k-th replication. The corresponding root-mean-square error (RMS) is given by :

$$\text{RMS} = \left\{\frac{1}{100}\sum_{k=1}^{100} (\hat{\lambda}(k) - \lambda)^2\right\}^{\frac{1}{2}}, \tag{2.21}$$

where λ is the true value of the parameter in second group.

By using $\hat{\lambda}$ and estimated standard error (S.E.) of $\hat{\lambda}$, we can make inference on testing of homogeneity on second group with reference in first group. If λ_0 is the true value of the parameter in first group, we are going to test $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$. The test statistic used is :

$$\chi^2 = \left\{ \frac{\hat{\lambda} - \lambda_0}{\hat{\text{S.E.}}(\hat{\lambda})} \right\}^2, \tag{2.22}$$

when the sample size is large. As χ^2 has an asymptotic chi-squared distribution with degree of freedom, $d.f.$=1. Then the p-value of the test (P-V) can be calculated. Besides, the following statistics are studied :

1. The average of estimated standard error of the parameter estimate,

$$\overline{\text{S.E.}} = \frac{1}{100} \sum_{k=1}^{100} \hat{\text{S.E.}}(\hat{\lambda}(k)), \tag{2.23}$$

   where $\hat{\text{S.E.}}(\hat{\lambda}(k))$ is the estimated standard error of $\hat{\lambda}(k)$ in the k-th replication.

2. The sample standard deviation of the parameter estimate,

$$\text{S.D.} = \left\{ \frac{1}{99} \sum_{k=1}^{100} (\hat{\lambda}(k) - \bar{\lambda})^2 \right\}^{\frac{1}{2}}. \tag{2.24}$$

3. The ratio of the sample standard deviation to the average of estimated standard error of the parameter estimate,

$$\text{RATIO} = \frac{\text{S.D.}}{\overline{\text{S.E.}}}. \tag{2.25}$$

Basically, S.D. and $\overline{\text{S.E.}}$ are both the estimates of the dispersion of $\hat{\lambda}$, they are expected to be closed. Therefore, RATIO should be close to 1.

After discussing all the above required statistics, we consider the study now. In the simulation study, various sets of threshold values and parameter value have been studied. The sample size of ranging from 100 to 1000 is studied in each simulation. There are mainly three cases considered, that is, cases (A), (B) and (C). In each case, there are also three sub-cases discussed according to the choice of the true thresholds. The first is called the symmetric distribution. It means that the thresholds are selected such that the grouped frequencies are almost equally divided in the distribution with the first group's parameter value. Explicitly, suppose considering the symmetric distribution with two thresholds, we thus assign $p_1 = 0.3$, $p_2 = 0.4$ and $p_3 = 0.3$ where p_i is the notation used in (2.4). The next one is called the asymmetric distribution. The thresholds are selected such that the shape of the distribution with the first group's parameter value is skewed to the right. Explicitly, we assign $p_1 = 0.6$, $p_2 = 0.3$ and $p_3 = 0.1$ in the asymmetric distribution with two thresholds. The last one is also called the asymmetric distribution but the shape of the distribution with the first group's parameter value is skewed to the left. Similarly, we assign $p_1 = 0.1$, $p_2 = 0.3$ and $p_3 = 0.6$ in the case of two thresholds. Different cases are discussed in the following settings :

(A) First group : $\lambda = 1.0$

Second group : $\lambda = 2.0$

A.1 Symmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.3567,\ 1.2040,\ +\infty)'.$$

A.2 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.9163,\ 2.3026,\ +\infty)'.$$

A.3 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.1054,\ 0.5108,\ +\infty)'.$$

(B) First group : $\lambda = 1.0$

Second group : $\lambda = 3.0$

B.1 Symmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.3567,\ 1.2040,\ +\infty)'.$$

B.2 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.9163,\ 2.3026,\ +\infty)'.$$

B.3 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.1054,\ 0.5108,\ +\infty)'.$$

(C) First group : $\lambda = 1.0$

Second group : $\lambda = 1.0$

C.1 Symmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.3567,\ 1.2040,\ +\infty)'.$$

C.2 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.9163,\ 2.3026,\ +\infty)'.$$

C.3 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.1054,\ 0.5108,\ +\infty)'.$$

In the following tables, some remarks should be noted :

1. Two numbers in the "SAMPLE SIZE" column are represented as : "1st group's sample size / 2nd group's sample size".

2. Only the maximum of the root-mean-square errors of the threshold estimates is presented in the "MAX(RMS_i) THRES." column.

3. In Cases (A) and (B), we consider two different values in the exponential parameter for the two groups. In doing the test:

$$H_0 : \lambda = \lambda_0 \qquad \text{vs.} \quad H_1 : \lambda \neq \lambda_0,$$

where λ_0 is the true value of the parameter in the first group, we consider the χ^2-statistic and compute the p-value of the test. In Tables 2.2 and 2.3, the number of the p-values greater than 0.05 is presented in the "NO. P-V>0.05" column. In Case (C), we consider the same value in the exponential parameter in two groups. Doing the same test as before, the number of the p-values smaller than 0.05 is presented in the "NO. P-V<0.05" column in Table 2.4.

The simulation study's results are :

Table 2.2: Simulation Study on Exponential Distribution. Case (A).

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
A.1	100/100	0.1422	2.0379	0.4115	3	0.4118	0.3101	1.3277
	200/200	0.1078	2.0563	0.2581	0	0.2531	0.2194	1.1536
	100/200	0.1436	2.0609	0.3098	0	0.3053	0.2199	1.3879
	200/100	0.1049	2.0123	0.3778	1	0.3795	0.3044	1.2467
A.2	100/100	0.3452	2.0663	0.3348	0	0.3299	0.2545	1.2962
	200/200	0.1999	1.9782	0.2085	0	0.2084	0.1716	1.2146
	100/200	0.2936	2.0290	0.2501	0	0.2497	0.1761	1.4182
	200/100	0.2088	2.0528	0.3257	0	0.3230	0.2547	1.2684
A.3	100/100	0.0750	2.0474	0.5754	26	0.5763	0.4481	1.2862
	200/200	0.0550	1.9905	0.3551	1	0.3568	0.2987	1.1945
	100/200	0.0814	2.0983	0.4220	4	0.4125	0.3181	1.2968
	200/100	0.0527	2.0538	0.5365	22	0.5365	0.4453	1.2047

Table 2.3: Simulation Study on Exponential Distribution. Case (B).

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
B.1	100/100	0.1617	3.1038	0.6946	0	0.6903	0.5529	1.2484
	200/200	0.1061	2.9877	0.3951	0	0.3969	0.3702	1.0721
	100/200	0.1531	3.1358	0.5637	0	0.5498	0.3911	1.4057
	200/100	0.1139	3.0713	0.6218	0	0.6208	0.5429	1.1435
B.2	100/100	0.3462	3.0877	0.5719	0	0.5679	0.4282	1.3265
	200/200	0.1999	3.0635	0.3891	0	0.3858	0.2994	1.2889
	100/200	0.2750	3.0170	0.4219	0	0.4237	0.2943	1.4396
	200/100	0.2099	3.0393	0.5345	0	0.5357	0.4208	1.2731
B.3	100/100	0.0873	3.4645	1.3394	3	1.2626	0.9498	1.3293
	200/200	0.0573	3.0314	0.6364	0	0.6388	0.5523	1.1565
	100/200	0.0802	3.1215	0.7560	0	0.7499	0.5683	1.3196
	200/100	0.0544	3.3648	1.1101	1	1.0537	0.9184	1.1474

Table 2.4: Simulation Study on Exponential Distribution. Case (C).

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V<0.05	S.D.	$\overline{S.E.}$	RATIO
C.1	100/100	0.1490	0.9977	0.1696	16	0.1704	0.1215	1.4024
	200/200	0.1061	1.0300	0.1253	13	0.1222	0.0891	1.3722
	400/400	0.0748	1.0131	0.0807	11	0.0800	0.0619	1.2933
	800/800	0.0482	0.9947	0.0524	13	0.0524	0.0428	1.2243
C.2	100/100	0.3329	1.0024	0.1515	14	0.1522	0.1113	1.3678
	200/200	0.2437	1.0216	0.1115	14	0.1099	0.0801	1.3722
	400/400	0.1416	0.9996	0.0748	13	0.0752	0.0553	1.3594
	800/800	0.1116	0.9993	0.0462	9	0.0465	0.0391	1.1883
C.3	100/100	0.0840	1.0152	0.2232	12	0.2238	0.1630	1.3735
	200/200	0.0581	1.0106	0.1550	11	0.1554	0.1147	1.3546
	400/400	0.0446	1.0127	0.1016	10	0.1013	0.0807	1.2545
	800/800	0.0265	0.9993	0.0712	10	0.0715	0.0502	1.2727

From the Tables 2.2 to 2.4, we observe that :

1. In most situations, the mean of the parameter estimate is close to the true value.

2. In most situations, RMS is small.

3. When the sample size increases,

   (a) the mean of the parameter estimate becomes closer to the true value generally.

   (b) RMS decreases.

   (c) the RATIO=S.D./$\overline{\text{S.E.}}$ would be improved and becomes closer to 1, especially in the case of symmetric distribution.

4. In Case (A), we consider the parameter values as 1.0 and 2.0 in first group and second group respectively. It is found that the number of p-values greater than 0.05 is quite large, especially in the Case (A.3). The sample size should be as large as 200 in both groups to keep the Type II error probability low. In Case (B), we consider the parameter values as 1.0 and 3.0 in first group and second group respectively. It is observed that almost all the numbers of the p-values greater than 0.05 are small. In this case, the sample size can be as small as 100 in both groups to keep the Type II error probability low.

5. In Case (C), we consider the same value in the exponential parameter in two groups. Doing the same test as before, the number of the p-values smaller than 0.05 is presented in Table 2.4. In both symmetric and asymmetric cases, this number is acceptable but not satisfactory for large sample sizes in both groups. Hence the Type I error probability is not low in this case.

2.2 Normal distribution

If Z is an observable random variable which is related to an underlying continuous variable X by the form in the (2.1) where $\alpha_1, \alpha_2, \ldots, \alpha_{s+1}$ are the thresholds of Z with $\alpha_1 = -\infty$ and $\alpha_{s+1} = +\infty$. Now, suppose X is normally distributed of the density function :

$$f(x) = f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

$$-\infty < x < +\infty, \ -\infty < \mu < +\infty, \ \sigma > 0. \qquad (2.26)$$

There are two parameters μ and σ involved in the normal distribution. They provide us with complete information about the exact location and scale of the distribution. Its mean is μ and variance is σ^2.

2.2.1 Estimation of Thresholds

In this part, the observed data and the parameters μ and σ are given, the thresholds $\alpha_2, \ldots, \alpha_s$ are estimated by maximum likelihood method. Similar to the exponential case, suppose a random sample of size n is observed and n_i is the observed frequency in the i-th cell for $i = 1, 2, \ldots, s$ with $\sum_{i=1}^{s} n_i = n$. The likelihood function is given by (2.3) and the probability of an observation falling in the i-th cell is followed by :

$$p_1 = \Pr(Z = 1) = \int_{\alpha_1}^{\alpha_2} f(x; \mu, \sigma)\, dx,$$

$$\vdots$$

$$p_s = \Pr(Z = s) = \int_{\alpha_s}^{\alpha_{s+1}} f(x; \mu, \sigma)\, dx. \tag{2.27}$$

The maximum likelihood estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ is the vector which maximizes the likelihood function $L(\boldsymbol{\alpha})$, or equivalently, minimizes the negative log-likelihood function :

$$F(\boldsymbol{\alpha}) = -\log\, L(\boldsymbol{\alpha}) \;\propto\; -\sum_{i=1}^{s} n_i\, \log\left\{\int_{\alpha_i}^{\alpha_{i+1}} g(x)\, dx\right\}, \tag{2.28}$$

where

$$g(x) = g(x; \mu, \sigma) = \frac{1}{\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{2.29}$$

To obtain the minimum of the negative log-likelihood function, the Newton-Raphson algorithm is applied. Therefore, the first and the second derivatives of the negative log-likelihood function $F(\boldsymbol{\alpha})$ are required to deduce. The results are given as follows :

$$\frac{\partial F(\boldsymbol{\alpha})}{\partial \alpha_j} = g(\alpha_j) \left\{\frac{n_j}{\int_{\alpha_j}^{\alpha_{j+1}} g(x)dx} - \frac{n_{j-1}}{\int_{\alpha_{j-1}}^{\alpha_j} g(x)dx}\right\} \qquad \text{for } j = 2, \ldots, s, \tag{2.30}$$

$$\frac{\partial^2 F(\boldsymbol{\alpha})}{\partial {\alpha_j}^2} = (g(\alpha_j))^2 \left\{\frac{n_{j-1}}{(\int_{\alpha_{j-1}}^{\alpha_j} g(x)dx)^2} + \frac{n_j}{(\int_{\alpha_j}^{\alpha_{j+1}} g(x)dx)^2}\right\} + \frac{(\mu - \alpha_j)}{\sigma^2} \frac{\partial F(\boldsymbol{\alpha})}{\partial \alpha_j}$$

for $j = 2, \ldots, s$. (2.31)

In addition, some of the non-zero entries of the Hessian are derived as :

$$\frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_{j-1} \partial \alpha_j} = -\frac{n_{j-1}\ g(\alpha_j)\ g(\alpha_{j-1})}{(\int_{\alpha_{j-1}}^{\alpha_j} g(x)dx)^2} \quad \text{for } j = 3, \ldots, s,$$

$$\frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_{j+1} \partial \alpha_j} = -\frac{n_j\ g(\alpha_j)\ g(\alpha_{j+1})}{(\int_{\alpha_j}^{\alpha_{j+1}} g(x)dx)^2} \quad \text{for } j = 2, \ldots, s-1. \tag{2.32}$$

Apart from the above entries, the remaining entries are all equal to zero in the Hessian matrix. The form of the Hessian matrix is given by (2.12).

2.2.2 Estimation of Parameter

In the second part, the thresholds are assumed to be the given constants, the parameters μ and σ are estimated also by maximum likelihood method. Suppose $\boldsymbol{\theta} = (\mu,\ \sigma)'$, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is one which maximizes the likelihood function $L(\boldsymbol{\theta})$, or equivalently, minimizes the negative log-likelihood function :

$$G(\boldsymbol{\theta}) = -\log\ L(\boldsymbol{\theta}) \ \propto\ -\sum_{i=1}^{s} n_i\ \log\left\{\int_{\alpha_i}^{\alpha_{i+1}} g(x;\mu,\sigma)\ dx\right\}, \tag{2.33}$$

where $g(x;\mu,\sigma)$ is the one shown in (2.29). By using the Newton-Raphson algorithm, the required derivatives of $G(\boldsymbol{\theta})$ are found as follows.

The first derivatives are :

$$\frac{\partial G(\boldsymbol{\theta})}{\partial \mu} = \sum_{i=1}^{s} \frac{n_i \, [g(\alpha_{i+1}) - g(\alpha_i)]}{\int_{\alpha_i}^{\alpha_{i+1}} g(x)dx},$$

$$\frac{\partial G(\boldsymbol{\theta})}{\partial \sigma} = \sum_{i=1}^{s} \frac{n_i \, [(\alpha_{i+1} - \mu)g(\alpha_{i+1}) - (\alpha_i - \mu)g(\alpha_i)]}{\sigma \int_{\alpha_i}^{\alpha_{i+1}} g(x)dx}. \tag{2.34}$$

For simplicity, letting

$q_i = \int_{\alpha_i}^{\alpha_{i+1}} g(x)\, dx,$

$h_1(x) = (x - \mu)g(x),$

$h_2(x) = (x - \mu)^2 g(x),$

$h_3(x) = (x - \mu)^3 g(x).$

Note: $\lim_{x \to -\infty} g(x) = 0, \ \lim_{x \to +\infty} g(x) = 0,$

$\lim_{x \to -\infty} h_i(x) = 0 \quad$ for $i = 1, 2, 3,$

$\lim_{x \to +\infty} h_i(x) = 0 \quad$ for $i = 1, 2, 3,$

and denote $[h_j(x)]_{\alpha_i}^{\alpha_{i+1}} = h_j(\alpha_{i+1}) - h_j(\alpha_i) \quad$ for $j = 1, 2, 3,$

also $\quad [g(x)]_{\alpha_i}^{\alpha_{i+1}} = g(\alpha_{i+1}) - g(\alpha_i).$

The second derivatives are :

$$\frac{\partial^2 G(\boldsymbol{\theta})}{\partial \mu^2} = \sum_{i=1}^{s} \frac{n_i \, [q_i \, [h_1(x)]_{\alpha_i}^{\alpha_{i+1}} + \sigma^2([g(x)]_{\alpha_i}^{\alpha_{i+1}})^2]}{\sigma^2 {q_i}^2},$$

$$\frac{\partial^2 G(\boldsymbol{\theta})}{\partial \sigma^2} = \sum_{i=1}^{s} \frac{n_i \, [q_i \, [h_3(x)]_{\alpha_i}^{\alpha_{i+1}} + \sigma^2([h_1(x)]_{\alpha_i}^{\alpha_{i+1}})^2 - 2\sigma^2 q_i \, [h_1(x)]_{\alpha_i}^{\alpha_{i+1}}]}{\sigma^4 {q_i}^2},$$

$$\begin{aligned}\frac{\partial^2 G(\boldsymbol{\theta})}{\partial\mu\partial\sigma} &= \frac{\partial^2 G(\boldsymbol{\theta})}{\partial\sigma\partial\mu} \\ &= \sum_{i=1}^{s} \frac{n_i\,[q_i\,[h_2(x)]_{\alpha_i}^{\alpha_{i+1}} + \sigma^2[g(x)]_{\alpha_i}^{\alpha_{i+1}}[h_1(x)]_{\alpha_i}^{\alpha_{i+1}} - \sigma^2 q_i\,[g(x)]_{\alpha_i}^{\alpha_{i+1}}]}{\sigma^3 {q_i}^2}.\end{aligned} \tag{2.35}$$

By using the Newton-Raphson algorithm,

$$\triangle\boldsymbol{\theta} = -\gamma H(\boldsymbol{\theta})^{-1}\dot{G}(\boldsymbol{\theta}), \tag{2.36}$$

where γ is a step-size parameter which may be taken as the first value in the sequence $1, \frac{1}{2}, \frac{1}{4}, \ldots$ that reduces G. $\dot{G}(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$ are the gradient vector and the Hessian matrix which can be found from (2.34) and (2.35). According to the asymptotic theory, since $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, it can be shown that,

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{l} N[0,\ \mathrm{I}^{-1}(\boldsymbol{\theta})]. \tag{2.37}$$

Actually, since the Hessian matrix converges in probability to $\mathrm{I}(\boldsymbol{\theta})$, the Hessian matrix can be used to approximate the information matrix. The maximum likelihood estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by $(H(\hat{\boldsymbol{\theta}}))^{-1}$. Therefore, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ can be obtained.

2.2.3 Simulation Study

Now, a simulation study on normal distribution is discussed. This study is based on simulating data from two normal distributions. The first one is taken as the reference group and the second is the one under the test of homogeneity. Similar to the exponential case, the thresholds are estimated in the reference group firstly. The mean of the threshold estimates and the root-mean-square error of the threshold estimates are obtained as given by (2.18) and (2.19) respectively. Based on the thresholds estimated, the value of the parameter in normal distribution can be estimated in second group's data. Since there are two parameters involved in normal distribution. The interesting statistics are double in number to that obtained in exponential case. All required statistics are analogous to those presented from (2.20) to (2.25). The followings are the cases considered :

(A) First group : $\mu = 0.0,\ \sigma = 1.0$

Second group : $\mu = 1.0,\ \sigma = 2.0$

A.1 Symmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (-\infty,\ -0.5244,\ 0.5244,\ +\infty)'.$$

A.2 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (-\infty,\ 0.2533,\ 1.2816,\ +\infty)'.$$

A.3 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (-\infty,\ -1.2816,\ -0.2533,\ +\infty)'.$$

(B) First group : $\mu = 0.0,\ \sigma = 1.0$

Second group : $\mu = 2.0,\ \sigma = 3.0$

B.1 Symmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (-\infty,\ -0.5244,\ 0.5244,\ +\infty)'.$$

B.2 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (-\infty,\ 0.2533,\ 1.2816,\ +\infty)'.$$

B.3 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (-\infty,\ -1.2816,\ -0.2533,\ +\infty)'.$$

(C) First group : $\mu = 0.0,\ \sigma = 1.0$

Second group : $\mu = 0.0,\ \sigma = 1.0$

C.1 Symmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (-\infty,\ -0.5244,\ 0.5244,\ +\infty)'.$$

C.2 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (-\infty,\ 0.2533,\ 1.2816,\ +\infty)'.$$

C.3 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (-\infty,\ -1.2816,\ -0.2533,\ +\infty)'.$$

Note :

1. Since there are two parameters involved in normal distribution, the means of the parameter estimates are presented in the sequence : $\hat{\mu}$, $\hat{\sigma}$ (in two separate rows), and similar in other related columns.

2. In the "P-V" column, our interests are not only $H_0 : \mu = \mu_0$, $H_0 : \sigma^2 = {\sigma_0}^2$, but also the simultaneous test on $H_0 : (\mu, \sigma^2)' = (\mu_0, {\sigma_0}^2)'$ where μ_0 and σ_0 are the true values of the parameters in first group.

The simulation study's results are :

Table 2.5: Simulation Study on Normal Distribution. Case (A).

CASE	SAMPLE SIZE	MAX(RMS$_i$) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
A.1	100/100	0.1461	0.9268	0.3703	7	0.3649	0.3156	1.1559
			1.9642	0.5262	15	0.5276	0.4523	1.1665
					3			
	200/200	0.0991	1.0395	0.2602	0	0.2585	0.2333	1.1080
			2.0447	0.3774	1	0.3767	0.3309	1.1382
					0			
	300/300	0.0778	0.9913	0.1685	0	0.1691	0.1796	0.9418
			1.9822	0.2856	0	0.2865	0.2544	1.1259
					0			
	100/300	0.1277	1.0162	0.2324	0	0.2330	0.1858	1.2539
			2.0020	0.3686	1	0.3704	0.2628	1.4096
					0			
	300/100	0.0780	1.0685	0.3770	1	0.3726	0.3568	1.0442
			2.1712	0.6294	7	0.6087	0.5174	1.1763
					1			

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
A.2	100/100	0.1802	1.0462	0.2895	4	0.2873	0.2456	1.1697
			2.0344	0.5032	11	0.5046	0.4221	1.1954
					0			
	200/200	0.1057	1.0163	0.1938	0	0.1941	0.1682	1.1535
			1.9999	0.3461	0	0.3479	0.2913	1.1942
					0			
	300/300	0.0924	1.0142	0.1463	0	0.1463	0.1380	1.0599
			2.0174	0.2806	0	0.2815	0.2379	1.1833
					0			
	100/300	0.1683	1.0112	0.1955	0	0.1962	0.1378	1.4244
			2.0130	0.3549	0	0.3564	0.2372	1.5027
					0			
	300/100	0.1115	1.0235	0.2392	0	0.2393	0.2448	0.9775
			2.0392	0.4611	4	0.4618	0.4260	1.0840
					0			

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{S.E.}$	RATIO
A.3	200/200	0.1348	1.0382	0.4033	18	0.4035	0.4426	0.9118
			2.0765	0.4624	11	0.4583	0.4779	0.9591
					18			
	300/300	0.1020	1.1081	0.3852	1	0.3716	0.3727	0.9972
			2.1187	0.3867	0	0.3699	0.3996	0.9256
					1			
	500/500	0.0915	1.0079	0.2615	0	0.2627	0.2712	0.9688
			2.0274	0.2849	0	0.2850	0.2914	0.9781
					0			
	200/500	0.1177	1.0482	0.3015	1	0.2991	0.2747	1.0888
			2.0561	0.3298	1	0.3267	0.2961	1.1032
					2			
	500/200	0.0832	1.0558	0.4234	19	0.4218	0.4672	0.9031
			2.1085	0.4843	14	0.4744	0.5048	0.9398
					16			

Table 2.6: Simulation Study on Normal Distribution. Case (B).

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
B.1	100/100	0.1273	2.3166	1.0807	1	1.0385	0.8664	1.1986
			3.3756	1.3153	3	1.2668	1.1246	1.1246
					0			
	150/150	0.1134	2.0204	0.6618	0	0.6649	0.5914	1.1241
			3.1261	0.9092	0	0.9049	0.7986	1.1332
					0			
	200/200	0.0945	2.1122	0.5738	0	0.5655	0.5357	1.0557
			3.2207	0.8255	0	0.7995	0.7193	1.1114
					0			
	100/200	0.1334	2.0684	0.6499	0	0.6495	0.5243	1.2389
			3.1457	0.8386	0	0.8300	0.6933	1.1973
					0			
	200/100	0.0939	2.2651	0.9711	1	0.9389	0.8578	1.0945
			3.3437	1.3053	1	1.2656	1.1288	1.1212
					0			

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
B.2	100/100	0.1946	2.2084	0.7230	0	0.6958	0.5938	1.1718
			3.4636	1.3508	1	1.2752	1.0314	1.2364
					0			
	150/150	0.1275	2.1678	0.5628	0	0.5399	0.4538	1.1898
			3.2646	0.9364	0	0.9028	0.7698	1.1728
					0			
	200/200	0.1299	2.0437	0.4258	0	0.4257	0.3591	1.1854
			3.1109	0.6905	0	0.6849	0.6155	1.1129
					0			
	100/200	0.1893	2.0775	0.5124	0	0.5091	0.3807	1.3374
			3.2195	0.9353	0	0.9138	0.6583	1.3881
					0			
	200/100	0.1114	2.0988	0.6078	0	0.6027	0.5254	1.1471
			3.0388	0.8554	1	0.8588	0.8645	0.9934
					0			

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
B.3	200/200	0.1197	2.1168	0.7581	2	0.7528	0.8405	0.8956
			3.1169	0.8339	1	0.8298	0.8770	0.9462
					1			
	300/300	0.1070	2.1754	0.6993	1	0.6803	0.6980	0.9747
			3.1216	0.7063	0	0.6993	0.7195	0.9719
					0			
	500/500	0.0705	1.9815	0.4771	0	0.4792	0.4808	0.9966
			2.9649	0.5041	0	0.5054	0.5035	1.0038
					0			
	200/500	0.1299	2.1459	0.5130	0	0.4943	0.5159	0.9582
			3.1505	0.5311	0	0.5118	0.5405	0.9470
					0			
	500/200	0.0849	2.1217	0.8118	4	0.8067	0.8411	0.9590
			3.1601	0.9200	1	0.9106	0.8848	1.0291
					1			

Table 2.7: Simulation Study on Normal Distribution. Case (C).

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V<0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
C.1	100/100	0.1307	-0.1238	0.1600	17	0.1603	0.1144	1.4014
			1.0179	0.1883	18	0.1884	0.1388	1.3566
					20			
	200/200	0.0929	-0.0122	0.1062	14	0.1060	0.0795	1.3333
			1.0060	0.1293	12	0.1298	0.0957	1.3556
					18			
	400/400	0.0723	-0.0012	0.0772	15	0.0776	0.0561	1.3832
			1.0055	0.0903	13	0.0906	0.0676	1.3401
					19			
	800/800	0.0517	-0.0007	0.0527	17	0.0529	0.0397	1.3344
			1.0062	0.0638	12	0.0639	0.0480	1.3299
					15			

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V<0.05	S.D.	$\overline{S.E.}$	RATIO
C.2	100/100	0.1858	-0.0143	0.2022	16	0.2027	0.1446	1.4021
			1.0037	0.2159	17	0.2170	0.1599	1.3570
					23			
	200/200	0.1324	-0.0065	0.1355	12	0.1359	0.1028	1.3231
			1.0165	0.1611	14	0.1610	0.1138	1.4140
					21			
	400/400	0.0929	0.0164	0.1019	15	0.1010	0.0699	1.4445
			0.9890	0.1079	16	0.1079	0.0772	1.3978
					21			
	800/800	0.0564	-0.0021	0.0667	17	0.0670	0.0501	1.3388
			0.9978	0.0740	10	0.0743	0.0554	1.3414
					20			

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V<0.05	S.D.	$\overline{S.E.}$	RATIO
C.3	400/400	0.0900	0.0079	0.1221	8	0.1225	0.1097	1.1165
			1.0114	0.1322	8	0.1323	0.1212	1.0915
					13			
	600/600	0.0737	0.0224	0.0901	3	0.0877	0.0952	0.9214
			1.0196	0.1008	5	0.0994	0.1049	0.9482
					12			
	800/800	0.0565	0.0077	0.0635	2	0.0634	0.0775	0.8177
			1.0124	0.0725	3	0.0718	0.0858	0.8373
					10			
	1000/1000	0.0543	-0.0143	0.0652	4	0.0639	0.0701	0.9112
			0.9907	0.0627	0	0.0624	0.0773	0.8066
					12			

From the Tables 2.5 to 2.7, we observe that :

1. In most situations, the mean of the parameter estimate is close to the true value.

2. In most situations, RMS is small.

3. The magnitude of RMS of σ is greater than that of μ. This is possibly due to the magnitude of σ is greater than that of μ.

4. When the sample size increases,

   (a) the mean of the parameter estimate becomes closer to the true value generally.

   (b) RMS decreases.

   (c) the RATIO=S.D./$\overline{\text{S.E.}}$ would be improved and becomes closer to 1.

5. In Cases (A) and (B), we consider two sets of different values in the normal parameter for the two groups. In doing the test:

$$H_0 : \mu = \mu_0, \quad H_0 : \sigma^2 = {\sigma_0}^2, \text{ and } H_0 : (\mu, \sigma^2)' = (\mu_0, {\sigma_0}^2)',$$

   where μ_0, σ_0 are the true values of the parameters in the first group, we thus compute the p-value of the each test. In Case (A), we consider the parameter values as $\mu = 0.0$, $\sigma = 1.0$ in first group and $\mu = 1.0$, $\sigma = 2.0$ in

second group. It is found that the number of p-values greater than 0.05 is quite large , especially in the Case (A.3). In this case, the sample size should be as large as 300 in both groups to keep the Type II error probability low. In Case (B), we consider the parameter values as $\mu = 0.0$, $\sigma = 1.0$ in first group and $\mu = 2.0$, $\sigma = 3.0$ in second group. It is observed that all the numbers of the p-values greater than 0.05 are smaller than 5. In this case, the sample size can be as small as 200 in both groups to keep the Type II error probability low.

6. In Case (C), we consider the same value in the normal parameter in two groups. Doing the same test as before, the number of the p-values smaller than 0.05 is presented in Table 2.7. However, this number is only acceptable but not satisfactory in both symmetric and asymmetric cases.

2.3 Weibull distribution

In this section, the variable X is supposed to follow a Weibull distribution. Similar to the previous sections, let Z be an observable random variable, relating to a continuous variable X by the form given in (2.1) where $\alpha_1, \alpha_2, \ldots, \alpha_{s+1}$ are the thresholds of Z with $\alpha_1 = 0$ and $\alpha_{s+1} = +\infty$. Suppose the variable X has a Weibull distribution of the density function :

$$f(x) = f(x; \gamma, \beta) = \frac{\gamma}{\beta} x^{\gamma-1} \exp\left\{-\frac{x^\gamma}{\beta}\right\}, \quad x \geq 0, \ \ \gamma, \beta > 0. \tag{2.38}$$

The Weibull distribution is determined by two parameters, γ and β where γ is the shape parameter and $\beta^{1/\gamma}$ is the scale parameter. It plays an important role in the analysis of failure time data.

2.3.1 Estimation of Thresholds

In this part, suppose a random sample of size n is observed, given the observed data and the parameters γ and β, the task is to estimate the thresholds $\alpha_2, \ldots, \alpha_s$ by maximum likelihood method. At first, the likelihood function considered is similar to the previous sections as given in (2.3) where p_i is the probability of an observation falling in the i-th cell. Then

$$p_1 = \Pr(Z = 1) = \int_{\alpha_1}^{\alpha_2} f(x; \gamma, \beta)\, dx,$$

$$\vdots$$

$$p_s = \Pr(Z = s) = \int_{\alpha_s}^{\alpha_{s+1}} f(x;\gamma,\beta)\, dx. \tag{2.39}$$

The maximum likelihood estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ is the vector which maximizes the likelihood function $L(\boldsymbol{\alpha})$, or equivalently, minimizes the negative log-likelihood function :

$$F(\boldsymbol{\alpha}) = -\log\, L(\boldsymbol{\alpha}) \;\propto\; -\sum_{i=1}^{s} n_i\, \log\left\{\int_{\alpha_i}^{\alpha_{i+1}} f(x)dx\right\}, \tag{2.40}$$

where

$$f(x) = f(x;\gamma,\beta) = \frac{\gamma}{\beta}x^{\gamma-1}\exp\left\{-\frac{x^{\gamma}}{\beta}\right\}. \tag{2.41}$$

Now, by using the Newton-Raphson algorithm to obtain the minimum of the negative log-likelihood function, the gradient vector and the Hessian matrix are required. The first and second derivatives are computed as follows :

$$\frac{\partial F(\boldsymbol{\alpha})}{\partial \alpha_j} = f(\alpha_j)\left\{\frac{n_j}{\int_{\alpha_j}^{\alpha_{j+1}} f(x)dx} - \frac{n_{j-1}}{\int_{\alpha_{j-1}}^{\alpha_j} f(x)dx}\right\} \qquad \text{for } j = 2,\ldots,s, \tag{2.42}$$

$$\begin{aligned}\frac{\partial^2 F(\boldsymbol{\alpha})}{\partial {\alpha_j}^2} \;=\; & (f(\alpha_j))^2\left\{\frac{n_{j-1}}{(\int_{\alpha_{j-1}}^{\alpha_j} f(x)dx)^2} + \frac{n_j}{(\int_{\alpha_j}^{\alpha_{j+1}} f(x)dx)^2}\right\} \\ & + \left\{\frac{\gamma\beta - \beta - \gamma{\alpha_j}^{\gamma}}{\beta\alpha_j}\right\}\frac{\partial F(\boldsymbol{\alpha})}{\partial \alpha_j} \qquad \text{for } j = 2,\ldots,s.\end{aligned} \tag{2.43}$$

The above second derivatives are the diagonal entries of the Hessian matrix.

Also, some of the off-diagonal entries are computed as:

$$\frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_{j-1} \partial \alpha_j} = - \frac{n_{j-1} f(\alpha_j) f(\alpha_{j-1})}{(\int_{\alpha_{j-1}}^{\alpha_j} f(x)dx)^2} \qquad \text{for } j = 3, \ldots, s,$$

$$\frac{\partial^2 F(\boldsymbol{\alpha})}{\partial \alpha_{j+1} \partial \alpha_j} = - \frac{n_j f(\alpha_j) f(\alpha_{j+1})}{(\int_{\alpha_j}^{\alpha_{j+1}} f(x)dx)^2} \qquad \text{for } j = 2, \ldots, s-1. \tag{2.44}$$

The entries other than the above are all equal to zero in the Hessian matrix.

2.3.2 Estimation of Parameter

In the second part, the thresholds are considered to be the fixed constants, the objective it to estimate the parameters γ and β in the Weibull distribution by maximum likelihood method. Suppose $\boldsymbol{\theta} = (\gamma,\ \beta)'$, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is one which maximizes the likelihood function $L(\boldsymbol{\theta})$, or equivalently, minimizes the negative log-likelihood function :

$$G(\boldsymbol{\theta}) = -\log\ L(\boldsymbol{\theta}) \ \propto\ -\sum_{i=1}^{s} n_i\ \log\left\{\int_{\alpha_i}^{\alpha_{i+1}} f(x;\gamma,\beta)\ dx\right\}, \tag{2.45}$$

where $f(x;\gamma,\beta)$ is given by (2.41). By using the Newton-Raphson algorithm, the required derivatives are found as follows.

For simplicity, letting

$h_1(x) = e^{-\frac{x^\gamma}{\beta}}$,

$h_2(x) = x^\gamma e^{-\frac{x^\gamma}{\beta}}$,

$h_3(x) = x^{2\gamma} e^{-\frac{x^\gamma}{\beta}}$,

$h_4(x) = x^\gamma e^{-\frac{x^\gamma}{\beta}} \log x$,

$h_5(x) = x^{2\gamma} e^{-\frac{x^\gamma}{\beta}} \log x$,

$h_6(x) = x^\gamma e^{-\frac{x^\gamma}{\beta}} (\log x)^2$,

$h_7(x) = x^{2\gamma} e^{-\frac{x^\gamma}{\beta}} (\log x)^2$.

Note: $\lim_{x\to 0} h_1(x) = 1$, $\lim_{x\to+\infty} h_1(x) = 0$,

$\lim_{x\to 0} h_i(x) = 0 \quad$ for $i = 2, 3, \ldots, 7$,

$\lim_{x\to+\infty} h_i(x) = 0 \quad$ for $i = 2, 3, \ldots, 7$,

and denote $[h_j(x)]_{\alpha_i}^{\alpha_{i+1}} = h_j(\alpha_{i+1}) - h_i(\alpha_i) \quad$ for $j = 1, 2, \ldots, 7$.

The first derivatives are :

$$\frac{\partial G(\boldsymbol{\theta})}{\partial \gamma} = \sum_{i=1}^{s} \frac{n_i \; [h_4(x)]_{\alpha_i}^{\alpha_{i+1}}}{\beta \; [h_1(x)]_{\alpha_i}^{\alpha_{i+1}}},$$

$$\frac{\partial G(\boldsymbol{\theta})}{\partial \beta} = \sum_{i=1}^{s} -\frac{n_i \; [h_2(x)]_{\alpha_i}^{\alpha_{i+1}}}{\beta^2 \; [h_1(x)]_{\alpha_i}^{\alpha_{i+1}}}. \tag{2.46}$$

The second derivatives are :

$$\frac{\partial^2 G(\boldsymbol{\theta})}{\partial \gamma^2} = \sum_{i=1}^{s} \frac{n_i}{\beta^2 \, [h_1(x)]_{\alpha_i}^{\alpha_{i+1}}} \left[\beta \, [h_6(x)]_{\alpha_i}^{\alpha_{i+1}} + \frac{([h_4(x)]_{\alpha_i}^{\alpha_{i+1}})^2}{[h_1(x)]_{\alpha_i}^{\alpha_{i+1}}} - [h_7(x)]_{\alpha_i}^{\alpha_{i+1}} \right],$$

$$\frac{\partial^2 G(\boldsymbol{\theta})}{\partial \beta^2} = \sum_{i=1}^{s} \frac{n_i}{\beta^4 \, [h_1(x)]_{\alpha_i}^{\alpha_{i+1}}} \left[2\beta \, [h_2(x)]_{\alpha_i}^{\alpha_{i+1}} + \frac{([h_2(x)]_{\alpha_i}^{\alpha_{i+1}})^2}{[h_1(x)]_{\alpha_i}^{\alpha_{i+1}}} - [h_3(x)]_{\alpha_i}^{\alpha_{i+1}} \right],$$

$$\begin{aligned}\frac{\partial^2 G(\boldsymbol{\theta})}{\partial\gamma\partial\beta} &= \frac{\partial^2 G(\boldsymbol{\theta})}{\partial\beta\partial\gamma} \\ &= \sum_{i=1}^{s} \frac{n_i}{\beta^3 \left[h_1(x)\right]_{\alpha_i}^{\alpha_{i+1}}} \left[\left[h_5(x)\right]_{\alpha_i}^{\alpha_{i+1}} - \beta \left[h_4(x)\right]_{\alpha_i}^{\alpha_{i+1}} - \frac{\left[h_2(x)\right]_{\alpha_i}^{\alpha_{i+1}} \left[h_4(x)\right]_{\alpha_i}^{\alpha_{i+1}}}{\left[h_1(x)\right]_{\alpha_i}^{\alpha_{i+1}}} \right].\end{aligned} \tag{2.47}$$

According to the asymptotic theory, since $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, it can be shown that,

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{l} N[0,\ \mathrm{I}^{-1}(\boldsymbol{\theta})]. \tag{2.48}$$

Actually, since the Hessian matrix converges in probability to $\mathrm{I}(\boldsymbol{\theta})$, the Hessian matrix can be used to approximate the information matrix. Therefore, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ can be obtained.

2.3.3 Simulation Study

In the last section of this chapter, a simulation study is discussed on Weibull distribution. From two groups of data, the estimation of thresholds and the subsequent estimation of parameters are performed in the reference group and the other group respectively. Under 100 replications, the required statistics are similar to that obtained in normal distribution as presented in § 2.2.3. The

followings are the cases considered :

(A) First group : $\gamma = 2.0,\ \beta = 6.0$

Second group : $\gamma = 1.0,\ \beta = 3.0$

A.1 Symmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 1.4629,\ 2.6877,\ +\infty)'.$$

A.2 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 2.3447,\ 3.7169,\ +\infty)'.$$

A.3 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.7951,\ 1.7507,\ +\infty)'.$$

(B) First group : $\gamma = 2.0,\ \beta = 6.0$

Second group : $\gamma = 0.5,\ \beta = 2.5$

B.1 Symmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 1.4629,\ 2.6877,\ +\infty)'.$$

B.2 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 2.3447,\ 3.7169,\ +\infty)'.$$

B.3 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.7951,\ 1.7507,\ +\infty)'.$$

(C) First group : $\gamma = 2.0,\ \beta = 6.0$

Second group : $\gamma = 2.0,\ \beta = 6.0$

C.1 Symmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 1.4629,\ 2.6877,\ +\infty)'.$$

C.2 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 2.3447,\ 3.7169,\ +\infty)'.$$

C.3 Asymmetric distribution with two thresholds,

$$\boldsymbol{\alpha} = (0,\ 0.7951,\ 1.7507,\ +\infty)'.$$

Note :

1. Since there are also two parameters involved in Weibull distribution, the means of the parameter estimates are presented in the sequence : $\hat{\gamma}$, $\hat{\beta}$ (in two separate rows), and similar in other related columns.

2. In the "P-V" column, our interests are not only $H_0 : \gamma = \gamma_0$, $H_0 : \beta = \beta_0$, but also the simultaneous test on $H_0 : (\gamma, \beta)' = (\gamma_0, \beta_0)'$ where γ_0 and β_0 are the true values of the parameters in first group.

The simulation study's results are :

Table 2.8: Simulation Study on Weibull Distribution. Case (A).

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{S.E.}$	RATIO
A.1	100/100	0.1659	1.0096	0.2609	4	0.2620	0.2064	1.2697
			3.2005	1.0009	17	0.9856	0.7339	1.3429
					5			
	200/200	0.1309	1.0129	0.1880	0	0.1885	0.1453	1.2969
			3.0987	0.6402	1	0.6357	0.4898	1.2980
					1			
	300/300	0.1033	0.9890	0.1310	0	0.1312	0.1174	1.1174
			2.9925	0.3886	0	0.3905	0.3776	1.0343
					0			
	150/300	0.1518	1.0244	0.1802	0	0.1794	0.1215	1.4769
			3.0899	0.5358	0	0.5309	0.4001	1.3266
					0			
	300/150	0.0953	1.0073	0.1810	0	0.1817	0.1701	1.0680
			3.1221	0.6356	4	0.6269	0.5742	1.0917
					1			

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{S.E.}$	RATIO
A.2	100/100	0.2944	0.9830	0.2677	8	0.2685	0.2209	1.2157
			3.1100	1.0590	26	1.0585	0.9234	1.1464
					3			
	200/200	0.1662	1.0219	0.1891	1	0.1857	0.1609	1.1731
			3.1976	0.7837	8	0.7622	0.6928	1.1328
					0			
	300/300	0.1477	1.0048	0.1354	0	0.1359	0.1312	1.0358
			3.0419	0.5427	3	0.5438	0.5138	1.0585
					0			
	150/300	0.2142	1.0096	0.1774	0	0.1780	0.1313	1.3558
			3.1325	0.7548	5	0.7469	0.5387	1.3865
					0			
	300/150	0.1437	1.0328	0.2108	3	0.2093	0.1895	1.1047
			3.2217	0.9678	15	0.9468	0.7939	1.1926
					0			

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{S.E.}$	RATIO
A.3	100/100	0.1589	1.0025	0.2739	5	0.2753	0.2115	1.3013
			3.0658	0.7853	7	0.7865	0.5696	1.3807
					2			
	200/200	0.0941	1.0038	0.1806	0	0.1815	0.1481	1.2253
			3.0841	0.4901	1	0.4853	0.3984	1.2179
					0			
	300/300	0.0850	1.0029	0.1474	0	0.1481	0.1219	1.2144
			3.1142	0.4061	0	0.3917	0.3287	1.1918
					0			
	150/300	0.1282	0.9793	0.1765	0	0.1761	0.1189	1.4818
			2.0561	0.3298	1	0.4522	0.3191	1.4171
					0			
	300/150	0.0843	1.0525	0.2258	2	0.2207	0.1765	1.2504
			3.1123	0.5486	1	0.5397	0.4669	1.1560
					2			

Table 2.9: Simulation Study on Weibull Distribution. Case (B).

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
B.1	100/100	0.1728	0.5180	0.1765	0	0.1765	0.1589	1.1104
			2.6436	0.6318	2	0.6179	0.5481	1.1275
					0			
	200/200	0.1208	0.5176	0.1193	0	0.1186	0.1116	1.0624
			2.5524	0.3736	0	0.3717	0.3630	1.0241
					0			
	300/300	0.1008	0.5073	0.0923	0	0.0925	0.0902	1.0250
			2.5532	0.2992	0	0.2959	0.2955	1.0012
					0			
	150/300	0.1504	0.5176	0.1177	0	0.1170	0.0915	1.2781
			2.5589	0.3317	0	0.3280	0.2973	1.1032
					0			
	300/150	0.0960	0.5161	0.1217	0	0.1212	0.1284	0.9445
			2.5697	0.4222	0	0.4185	0.4223	0.9911
					0			

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
B.2	100/100	0.2609	0.5037	0.2039	0	0.2049	0.1698	1.2072
			2.6197	0.7664	7	0.7608	0.6782	1.1218
					0			
	200/200	0.1825	0.4985	0.1361	0	0.1367	0.1212	1.1283
			2.5472	0.4905	1	0.4907	0.4940	1.0809
					0			
	300/300	0.1479	0.4939	0.1044	0	0.1048	0.0989	1.0590
			2.5067	0.3844	0	0.3863	0.3606	1.0712
					0			
	150/300	0.1978	0.5031	0.1133	0	0.1138	0.0994	1.1441
			2.5572	0.4582	1	0.4569	0.3738	1.2225
					0			
	300/150	0.1338	0.4954	0.1560	0	0.1568	0.1391	1.1273
			2.5777	0.5906	3	0.5884	0.5332	1.1036
					0			

CASE	SAMPLE SIZE	MAX(RMS$_i$) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V>0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
B.3	100/100	0.1543	0.5275	0.1858	0	0.1847	0.1527	1.2095
			2.6400	0.5533	2	0.5380	0.4657	1.1552
					0			
	200/200	0.0994	0.5076	0.1155	0	0.1158	0.1034	1.1194
			2.5126	0.3316	0	0.3329	0.3022	1.1018
					0			
	300/300	0.0794	0.4947	0.0933	0	0.0936	0.0828	1.1308
			2.5074	0.2520	0	0.2532	0.2453	1.0321
					0			
	150/300	0.1150	0.5175	0.1117	0	0.1109	0.0874	1.2693
			2.5099	0.2524	0	0.2535	0.2471	1.0257
					0			
	300/150	0.0877	0.5160	0.1391	0	0.1388	0.1208	1.1495
			2.5483	0.3827	0	0.3816	0.3569	1.0691
					0			

Table 2.10: Simulation Study on Weibull Distribution. Case (C).

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V<0.05	S.D.	$\overline{S.E.}$	RATIO
C.1	200/200	0.1293	2.0339	0.2764	17	0.2757	0.2031	1.3577
			6.4182	1.8612	14	1.8227	1.2943	1.4082
					26			
	400/400	0.0831	2.0311	0.1924	16	0.1908	0.1445	1.3212
			6.2960	1.2092	9	1.1784	0.8874	1.3279
					19			
	600/600	0.0781	2.0094	0.1538	11	0.1543	0.1166	1.3230
			6.0956	0.8594	11	0.8584	0.6895	1.2449
					20			
	800/800	0.0643	2.0079	0.1254	13	0.1257	0.1009	1.2464
			6.0716	0.7272	15	0.7273	0.5928	1.2268
					19			

CASE	SAMPLE SIZE	MAX(RMS_i) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V<0.05	S.D.	$\overline{S.E.}$	RATIO
C.2	200/200	0.1657	2.0338	0.3213	16	0.3211	0.2222	1.4444
			6.6046	2.6186	14	2.5607	1.7487	1.4644
					31			
	400/400	0.1357	1.9577	0.1999	12	0.1964	0.1539	1.2760
			5.8535	1.3165	19	1.3149	1.0522	1.2498
					25			
	600/600	0.1003	1.9933	0.1637	12	0.1644	0.1270	1.2947
			6.0591	1.0919	11	1.0958	0.8914	1.2292
					24			
	800/800	0.0786	2.0029	0.1428	13	0.1434	0.1104	1.3000
			6.0831	1.0281	14	1.0298	0.7768	1.3258
					19			

CASE	SAMPLE SIZE	MAX(RMS$_i$) THRES.	MEAN PAR EST.	RMS PAR EST.	NO. P-V<0.05	S.D.	$\overline{\text{S.E.}}$	RATIO
C.3	200/200	0.1044	1.9845	0.3322	13	0.3335	0.2607	1.2795
			6.2061	1.4962	14	1.4894	1.1303	1.3177
					22			
	400/400	0.0643	2.0137	0.2448	18	0.2456	0.1855	1.3242
			6.0348	0.8813	13	0.8851	0.7545	1.1729
					20			
	600/600	0.0528	2.0247	0.2142	14	0.2139	0.1522	1.4054
			6.0825	0.8237	12	0.8237	0.6230	1.3222
					18			
	800/800	0.0467	2.0055	0.1816	15	0.1825	0.1310	1.3929
			6.1408	0.7154	7	0.7050	0.5461	1.2908
					17			

From the Tables 2.8 to 2.10, we observe that :

1. In most situations, the mean of the parameter estimate is close to the true value.

2. In most situations, RMS is quite small.

3. The magnitude of RMS of β is greater than that of γ. This is possibly due to the magnitude of β is greater than that of γ.

4. When the sample size increases,

   (a) the mean of the parameter estimate becomes closer to the true value generally.

   (b) RMS decreases.

   (c) the RATIO=S.D./$\overline{\text{S.E.}}$ would be improved and becomes closer to 1.

5. In Cases (A) and (B), we consider two sets of different values in the Weibull parameter for the two groups. In doing the test:

$$H_0 : \gamma = \gamma_0, \quad H_0 : \beta = \beta_0, \text{ and } H_0 : (\gamma, \beta)' = (\gamma_0, \beta_0)',$$

   where γ_0, β_0 are the true values of the parameters in the first group, we thus compute the p-value of the each test. In Case (A), we consider the parameter values as $\gamma = 2.0$, $\beta = 6.0$ in first group and $\gamma = 1.0$, $\beta = 3.0$

in second group. It is observed that the number of p-values greater than 0.05 is quite large in both symmetric and asymmetric cases. In this case, the sample size should be as large as 300 in both groups to keep the Type II error probability low. In Case (B), we consider the parameter values as $\gamma = 2.0,\ \beta = 6.0$ in first group and $\gamma = 0.5,\ \beta = 2.5$ in second group. It is observed that almost all the numbers of the p-values greater than 0.05 are smaller than 5. In this case, the sample size can be as small as 200 in both groups to keep the Type II error probability low.

6. In Case (C), we consider the same value in the Weibull parameter in two groups. Doing the same test as before, the number of the p-values smaller than 0.05 is presented in Table 2.10. Similar to previous cases, this number is only acceptable but not satisfactory even for large sample sizes in both groups.

Chapter 3

Goodness-of-fit Test

In Chapter 2, we have considered three distributions, that is, exponential distribution, normal distribution and Weibull distribution. At first, for each distribution, given the observed data and the value of the parameter, the thresholds can be estimated by maximum likelihood method. Secondly, for the same distribution picked before, given the thresholds estimated (by assuming the thresholds are the given constants), we can estimate the parameter in the same distribution.

Considering the data given in Table 2.1, it is necessary to impose a univariate distribution on the reference year, 1972, say. Next, we can study the distribution on each year 1973 to year 1975 relatively by estimating the value of the parameter in the same distribution considered. However, given a data set, the problem is how to choose a suitable one from the three distributions currently

discussed or other possible distributions. Therefore, the goodness-of-fit test is considered in this chapter.

On the reference group, it is required to estimate the thresholds given the value of the parameter in the distribution chosen. Suppose a random sample of size n is observed and the threshold parameter is $\boldsymbol{\alpha} = \{\alpha_2, \ldots, \alpha_s\}$ with known α_1 and α_{s+1}. Let n_i be the observed frequency in each i-th cell for $i = 1, 2, \ldots, s$ with $\sum_{i=1}^{s} n_i = n$ and p_i be the corresponding probability of an observation falling in the i-th cell with $\sum_{i=1}^{s} p_i = 1$. This is a multinomial model with s cells. Then the likelihood function is given by :

$$L(\boldsymbol{\alpha}) = \frac{n!}{\prod_{i=1}^{s} n_i!} \prod_{i=1}^{s} p_i(\boldsymbol{\alpha})^{n_i}. \tag{3.1}$$

To test whether this chosen distribution fits the reference group's data, consider testing :

H_0: The distribution fits the data in the reference group,

vs. H_1: not H_0.

The likelihood ratio test statistic is defined as :

$$\begin{aligned} \lambda^* &= \frac{\max_{H_0} L(\boldsymbol{\alpha})}{\max_{H_0 \cup H_1} L(\boldsymbol{\alpha})} \\ &= \frac{\prod_{i=1}^{s} \hat{p}_i(\boldsymbol{\alpha})^{n_i}}{\prod_{i=1}^{s} \tilde{p}_i(\boldsymbol{\alpha})^{n_i}}. \end{aligned} \tag{3.2}$$

Under $H_0 \cup H_1$, there is only one constraint, $\sum_{i=1}^{s} p_i = 1$. The maximum likelihood

estimate of p_i is n_i/n for $i = 1, 2, \ldots, s$. Therefore,

$$\tilde{p}_i(\boldsymbol{\alpha}) = \frac{n_i}{n} \qquad \text{for } i = 1, 2, \ldots, s. \tag{3.3}$$

Under H_0, it is supposed that the chosen distribution fits the data. Given the value of the parameter in the distribution, the only thing to do is to estimate the threshold parameter, $\boldsymbol{\alpha}$. Therefore,

$$\begin{aligned}
\hat{p}_1(\boldsymbol{\alpha}) &= \int_{\alpha_1}^{\hat{\alpha_2}} f(x;\boldsymbol{\theta})\, dx, \\
\hat{p}_i(\boldsymbol{\alpha}) &= \int_{\hat{\alpha_i}}^{\alpha_i\hat{+}1} f(x;\boldsymbol{\theta})\, dx \qquad \text{for } i = 2, 3, \ldots, s-1, \\
\hat{p}_s(\boldsymbol{\alpha}) &= \int_{\hat{\alpha_s}}^{\alpha_{s+1}} f(x;\boldsymbol{\theta})\, dx,
\end{aligned} \tag{3.4}$$

where $f(x;\boldsymbol{\theta})$ is the density function of the distribution considered with $\boldsymbol{\theta}$ being the parameter vector involved and $\hat{\boldsymbol{\alpha}} = \{\hat{\alpha_2}, \ldots, \hat{\alpha_s}\}$ is the maximum likelihood estimate of $\boldsymbol{\alpha}$. Under H_0, it can be shown that,

$$G^2 = -2\log\lambda^* \xrightarrow{l} \chi^2_{d.f.}, \tag{3.5}$$

when the sample size is large. The number of degrees of freedom is given by :

$$\begin{aligned}
d.f. &= \text{number of parameters estimated under } H_0 \cup H_1 \\
&\quad - \text{number of parameters estimated under } H_0 \\
&= (s-1) - (s-1) \\
&= 0.
\end{aligned}$$

This means that this distribution can perfectly fit the data in the reference group. In other words, the reference group can be perfectly fitted by any distribution no matter how many parameters are involved in the distribution.

Once the thresholds can be estimated in the reference group, the following task is to estimate the value of the parameter in the same distribution imposed on the reference group in the remaining group's data. However, it is also required to perform the goodness-of-fit test in following sections.

3.1 Test for the Exponential distribution

Just before, we assume that a random sample of size n is observed and the number of thresholds is equal to $s-1$. Therefore, the vector of thresholds is $\boldsymbol{\alpha} = \{\alpha_2, \ldots, \alpha_s\}$ with known $\alpha_1 = 0$ and $\alpha_{s+1} = +\infty$. In this context, assume the thresholds are the given fixed constants. Out of the total sample, suppose n_i is the observed frequency in each i-th cell for $i = 1, 2, \ldots, s$ with $\sum_{i=1}^{s} n_i = n$. Also, let p_i be the probability that each observation falls in the i-th cell with $\sum_{i=1}^{s} p_i = 1$. Note that it is a multinomial model with s cells, then the likelihood function is :

$$L(p_i) = \frac{n!}{\prod_{i=1}^{s} n_i!} \prod_{i=1}^{s} p_i^{n_i}. \tag{3.6}$$

To test whether exponential distribution fits the data, consider the hypothesis as :

$$H_0 : Exp(\lambda) \text{ fits the data} \quad \text{vs.} \quad H_1 : \text{not } H_0.$$

The likelihood ratio test statistic is defined as :

$$\begin{aligned} \lambda^* &= \frac{\max_{H_0} L(p_i)}{\max_{H_0 \cup H_1} L(p_i)} \\ &= \frac{\max_{\lambda} L(p_i)}{\max_{p_i} L(p_i)} \\ &= \frac{\max_{\lambda} \prod_{i=1}^{s} p_i^{n_i}}{\max_{p_i} \prod_{i=1}^{s} p_i^{n_i}} \\ &= \frac{\prod_{i=1}^{s} \hat{p}_i^{n_i}}{\prod_{i=1}^{s} \tilde{p}_i^{n_i}}. \end{aligned} \tag{3.7}$$

Under $H_0 \cup H_1$, the only one constraint is $\sum_{i=1}^{s} p_i = 1$. The maximum likelihood estimate of p_i can be shown as n_i/n for $i = 1, 2, \ldots, s$. So,

$$\tilde{p}_i = \frac{n_i}{n} \qquad \text{for } i = 1, 2, \ldots, s, \tag{3.8}$$

and $\hat{p}_i$ is estimated by

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x; \hat{\lambda})\, dx \qquad \text{for } i = 1, 2, \ldots, s, \tag{3.9}$$

where

$$f(x; \lambda) = \frac{1}{\lambda} \exp\left\{-\frac{x}{\lambda}\right\}, \quad x \geq 0, \quad \lambda > 0, \tag{3.10}$$

and $\hat{\lambda}$ is the maximum likelihood estimate of λ which is deduced in § 2.1.2. Under H_0, it can be shown that,

$$G^2 = -2 \log \lambda^* \xrightarrow{l} \chi^2_{d.f.}, \tag{3.11}$$

when the sample size is large. The number of degrees of freedom is given by :

$$\begin{aligned} d.f. &= \text{number of parameters estimated under } H_0 \cup H_1 \\ &\quad - \text{number of parameters estimated under } H_0 \\ &= (s-1)-1 \\ &= s-2. \end{aligned}$$

Therefore, we reject H_0 if $G^2 \geq \chi^2{}_{d.f.,\alpha}$ where $\chi^2{}_{d.f.,\alpha}$ is the upper α point of the chi-squared distribution with $d.f. = s-2$. Depending on the number of cells, the number of degrees of freedom may be negative, zero or positive. Then, we come to a conclusion that three situations may happen :

1. If $s < 2$, it is an unidentified case.

2. If $s = 2$, the number of degrees of freedom is equal to zero. This means the exponential distribution fits the data perfectly.

3. If $s > 2$, the number of degrees of freedom must be an integer greater than zero. H_0 may or may not be rejected at certain level of significance. Then the exponential distribution may or may not fit the data.

3.2 Test for the Normal distribution

Suppose a random sample of size n is observed and there are $s-1$ thresholds, $\boldsymbol{\alpha} = \{\alpha_2, \ldots, \alpha_s\}$, say, with known $\alpha_1 = -\infty$ and $\alpha_{s+1} = +\infty$. Let n_i be the observed frequency falling in the i-th cell for $i = 1, 2, \ldots, s$ with $\sum_{i=1}^{s} n_i = n$ and also, p_i be the probability of an observation falling in the i-th cell for $i = 1, 2, \ldots, s$ with $\sum_{i=1}^{s} p_i = 1$. This leads to the multinomial model with s cells. Now, we assume the thresholds are the fixed constants, the likelihood function is the same as (3.6). Under the general hypothesis $H_0 \cup H_1$, there is only one constraint $\sum_{i=1}^{s} p_i = 1$. The maximum likelihood estimate of p_i can be found as n_i/n for $i = 1, 2, \ldots, s$. In this section, we are going to test whether normal distribution fits the observed data. Consider testing :

$$H_0 : N(\mu, \sigma^2) \text{ fits the data} \quad \text{vs.} \quad H_1 : \text{not } H_0.$$

The likelihood ratio test statistic is defined as :

$$\begin{aligned} \lambda^* &= \frac{\max_{H_0} L(p_i)}{\max_{H_0 \cup H_1} L(p_i)} \\ &= \frac{\max_{\mu,\sigma} L(p_i)}{\max_{p_i} L(p_i)} \\ &= \frac{\max_{\mu,\sigma} \prod_{i=1}^{s} p_i^{n_i}}{\max_{p_i} \prod_{i=1}^{s} p_i^{n_i}} \\ &= \frac{\prod_{i=1}^{s} \hat{p}_i^{n_i}}{\prod_{i=1}^{s} \tilde{p}_i^{n_i}}, \end{aligned} \tag{3.12}$$

where $\tilde{p}_i$ is given in (3.8) and $\hat{p}_i$ is estimated by

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x;\hat{\mu},\hat{\sigma})\,dx \qquad \text{for } i = 1,2,\ldots,s, \tag{3.13}$$

where

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$
$$-\infty < x < +\infty,\ -\infty < \mu < +\infty,\ \sigma > 0, \tag{3.14}$$

and $(\hat{\mu},\hat{\sigma})'$ is the maximum likelihood estimate of $(\mu,\sigma)'$. When the sample size is large, under H_0, it can be shown that (3.11) holds while the number of degrees of freedom is given by :

$$\begin{aligned} d.f. &= \text{number of parameters estimated under } H_0 \cup H_1 \\ &\quad - \text{number of parameters estimated under } H_0 \\ &= (s-1)-2 \\ &= s-3. \end{aligned}$$

Then we can reject H_0 if $G^2 \geq \chi^2_{d.f.,\alpha}$ where $\chi^2_{d.f.,\alpha}$ is the upper α point of the chi-squared distribution with $d.f. = s-3$. From the above result, it can be noticed that if the number of cells, $s = 3$, the number of degrees of freedom is equal to zero. This is the case of perfect fit. This means that if the number of thresholds considered are only two, $\boldsymbol{\alpha} = \{\alpha_2, \alpha_3\}$, normal distribution may perfect fit the data. However, if $s < 3$, the number of degrees of freedom is

negative. This is an unidentified case. If $s > 3$, the number of degrees of freedom is greater than zero. This means that the normal distribution may or may not fit the data well.

3.3 Test for the Weibull distribution

Similar to the previous sections, we are going to construct a likelihood ratio test statistic to test whether Weibull distribution fits the data. Consider a random sample of size n is observed and assume the thresholds $\alpha_2, \alpha_3, \ldots, \alpha_s$ are the fixed constants while α_1=0 and $\alpha_{s+1} = +\infty$ are known. Let n_i be the observed frequency falling in the i-th cell for $i = 1, 2, \ldots, s$ with $\sum_{i=1}^{s} n_i = n$ and p_i be the probability of an observation falling in the i-th cell for $i = 1, 2, \ldots, s$ with $\sum_{i=1}^{s} p_i = 1$. Since it is a multinomial model with s cells, the likelihood function is also given by (3.6). Now, we are going to test :

$$H_0 : Wei(\gamma, \beta) \text{ fits the data} \quad \text{vs.} \quad H_1 : \text{not } H_0.$$

The likelihood ratio test statistic is defined as :

$$\begin{aligned} \lambda^* &= \frac{\max_{H_0} L(p_i)}{\max_{H_0 \cup H_1} L(p_i)} \\ &= \frac{\max_{\gamma,\beta} L(p_i)}{\max_{p_i} L(p_i)} \\ &= \frac{\max_{\gamma,\beta} \prod_{i=1}^{s} p_i^{n_i}}{\max_{p_i} \prod_{i=1}^{s} p_i^{n_i}} \\ &= \frac{\prod_{i=1}^{s} \hat{p}_i^{\,n_i}}{\prod_{i=1}^{s} \tilde{p}_i^{\,n_i}}. \end{aligned} \tag{3.15}$$

Under $H_0 \cup H_1$, there is only one constraint $\sum_{i=1}^{s} p_i = 1$. The maximum likelihood estimate of p_i is n_i/n for $i = 1, 2, \ldots, s$. In the numerator of the test statistic, $\hat{p}_i$ is estimated by

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x; \hat{\gamma}, \hat{\beta})\, dx \qquad \text{for } i = 1, 2, \ldots, s, \tag{3.16}$$

where

$$f(x; \gamma, \beta) = \frac{\gamma}{\beta} x^{\gamma-1} \exp\left\{-\frac{x^\gamma}{\beta}\right\}, \quad x \geq 0,\ \gamma, \beta > 0, \tag{3.17}$$

and $(\hat{\gamma}, \hat{\beta})'$ is the maximum likelihood estimate of $(\gamma, \beta)'$. When the sample size is large, under H_0, it can also be shown that (3.11) holds where the number of degrees of freedom is given by :

$$\begin{aligned} d.f. &= \text{number of parameters estimated under } H_0 \cup H_1 \\ &\quad - \text{number of parameters estimated under } H_0 \\ &= (s-1) - 2 \\ &= s - 3. \end{aligned}$$

Therefore, H_0 is rejected if $G^2 \geq \chi^2_{d.f.,\alpha}$ where $\chi^2_{d.f.,\alpha}$ is the upper α point of the chi-squared distribution with $d.f. = s - 3$. Actually, it is the case which is similar to the normal distribution case. There are three situations considered :

1. If $s = 3$, the number of degrees of freedom is equal to zero. This means that the Weibull distribution can perfect fit the data.

2. If $s < 3$, the number of degrees of freedom is negative. It is an unidentified case.

3. If $s > 3$, the number of degrees of freedom is positive. It means that the Weibull distribution may or may not fit the data.

3.4 Implication

Based on the above three distributions, the likelihood ratio test statistics can be constructed. In the exponential distribution, there is only one parameter involved. In the normal or the Weibull distribution, two parameters are involved. In deriving the likelihood ratio test statistic, it is observed that the number of degrees of freedom depends on the number of the thresholds and the number of the parameters involved in the distribution. Therefore, we can generalize the result and consider a family of distribution determined by k parameters, $q_1, q_2, \ldots, q_k$, say. Following with this chapter's setting, assume that there are $s-1$ thresholds, $\alpha_2, \alpha_3, \ldots, \alpha_s$ with known α_1 and α_{s+1}. Suppose n is the total observed sample size and n_i is the observed frequency falling in the i-th cell with the corresponding probability p_i for $i = 1, 2, \ldots, s$. In testing whether this distribution fits the data, consider testing :

$$H_0 : \text{ a family of distribution fits the data} \quad \text{vs.} \quad H_1 : \text{not } H_0.$$

The likelihood ratio test statistic is defined by :

$$\begin{aligned} \lambda^* &= \frac{\max_{parameters\ involved} L(p_i)}{\max_{p_i} L(p_i)} \\ &= \frac{\max_{parameters\ involved} \prod_{i=1}^{s} p_i^{n_i}}{\max_{p_i} \prod_{i=1}^{s} p_i^{n_i}} \\ &= \frac{\prod_{i=1}^{s} \hat{p}_i^{\,n_i}}{\prod_{i=1}^{s} \tilde{p}_i^{\,n_i}}. \end{aligned} \tag{3.18}$$

In the denominator, $\tilde{p}_i = n_i/n \qquad$ for $i = 1, \ldots, s$.

In the numerator, $\hat{p}_i$ is estimated by

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x; \hat{q}_1, \ldots, \hat{q}_k)\, dx \qquad \text{for } i = 1, 2, \ldots, s, \tag{3.19}$$

where $f(x; q_1, \ldots, q_k)$ is the density function of the distribution considered and $(\hat{q}_1, \ldots, \hat{q}_k)'$ is the maximum likelihood estimate of $(q_1, \ldots, q_k)'$. When the sample size is large, under H_0, it can be shown that,

$$G^2 = -2 \log \lambda^* \xrightarrow{l} \chi^2_{d.f.}, \tag{3.20}$$

where the number of degrees of freedom is given by :

$$\begin{aligned} d.f. &= \text{number of parameters estimated under } H_0 \cup H_1 \\ &\quad - \text{number of parameters estimated under } H_0 \\ &= (s-1) - \text{number of independent parameters estimated} \\ &= (s-1) - k. \end{aligned}$$

Therefore, we come to a conclusion that three cases may happen :

1. If $k > s - 1$, it is an unidentified case.

2. If $k = s - 1$, the number of degrees of freedom is equal to zero. It is a case of perfect fit. This family of distribution can fit the data well.

3. If $k < s - 1$, the number of degrees of freedom is greater than zero. Then this family of distribution may or may not fit the data well.

3.5 An Artificial Example

To illustrate the goodness-of-fit test, we consider the following example. Suppose there are three groups of data and each group has a sample size of 1000. Group 1 is generated from $N(0,1)$, group 2 is generated from $N(0,2)$ and group 3 is generated from $N(1,2)$. Two cases are discussed as follows :

3.5.1 Case 1 ($s = 3$)

Suppose the number of thresholds is equal to 2, that is, $\boldsymbol{\alpha} = \{\alpha_2, \alpha_3\}$ with known α_1 and α_4. Assume that the true values of thresholds α_2 and α_3 are equal to -2 and $+2$ respectively. The data generated are presented in Table 3.1. In this setting, we take group 1 as the reference group. We are going to discuss the goodness-of-fit test in the three distributions.

Table 3.1: Artificial 3×3 Contingency Table.

	Category			
	1	2	3	Total
Group 1	22	953	25	1000
Group 2	150	701	149	1000
Group 3	77	632	291	1000

3.5.1.1 Exponential distribution

In this section, we impose an exponential distribution on group 1. At first, the value of the parameter λ which determines the exponential distribution is assumed to one. By maximum likelihood method, the thresholds α_2 and α_3 can be estimated. The result is :

$$\hat{\alpha_2} = 0.0222, \quad \hat{\alpha_3} = 3.6889.$$

Next, by assuming the thresholds found are the fixed constants, we can estimate the value of the parameter in the exponential distribution on group 2 and group 3 by maximum likelihood method. The result is :

$$\hat{\lambda}_{gp.2} = 1.6612, \quad \hat{\lambda}_{gp.3} = 2.8254.$$

To test whether exponential distribution fits the data, consider the hypothesis as

:

$$H_0 : \text{The data fits the reference distribution} \quad \text{vs.} \quad H_1 : \text{not } H_0.$$

In group 2, the likelihood ratio test statistic is given by (3.7),

$$\lambda^* = \frac{\prod_{i=1}^{3} \hat{p}_i^{n_i}}{\prod_{i=1}^{3} \tilde{p}_i^{n_i}},$$

where

$$\tilde{p}_i = \frac{n_i}{n} \quad \text{for } i = 1, 2, 3,$$

and $\hat{p}_i$ is estimated by

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x; \hat{\lambda}_{gp.2})\, dx \quad \text{for } i = 1, 2, 3.$$

Under H_0,

$$G^2 = -2 \log \lambda^* \xrightarrow{l} {\chi^2}_{d.f.},$$

where the number of degrees of freedom, $d.f. = s - 2 = 1$. The calculated G^2 is 505.33 which is much greater than ${\chi^2}_{1,0.05} = 3.841$. Therefore, H_0 is rejected at $\alpha = 0.05$ level of significance. In group 3, the likelihood ratio test statistic λ^* can also be found as the above procedure. Under H_0,

$$G^2 = -2 \log \lambda^* \xrightarrow{l} {\chi^2}_{d.f.},$$

where $d.f. = s - 2 = 1$. The calculated G^2 is 226.40, also much greater than ${\chi^2}_{1,0.05} = 3.841$. Therefore, H_0 is rejected at $\alpha = 0.05$.

As an objective comparison, we assume the value of the parameter λ which determines the exponential distribution on group 1 as two. By maximum likelihood method, the thresholds α_2 and α_3 are estimated as :

$$\hat{\alpha_2} = 0.0445, \quad \hat{\alpha_3} = 7.3778.$$

Next, by assuming the thresholds found are the fixed constants, we can estimate the value of the parameter in the exponential distribution on group 2 and group 3 by maximum likelihood method. The result is :

$$\hat{\lambda}_{gp.2} = 3.3225, \quad \hat{\lambda}_{gp.3} = 5.6507.$$

Then the likelihood ratio test statistic, λ^* for testing whether exponential fits the data can be calculated. Same as the above setting, under H_0,

$$G^2 = -2\log\lambda^* \xrightarrow{l} \chi^2_{d.f.},$$

where the number of degrees of freedom, $d.f. = s - 2 = 1$. The calculated G^2 are 505.33 and 226.40 in group 2 and group 3 respectively. Both are much greater than $\chi^2_{1,0.05} = 3.841$. Therefore, H_0 are rejected at $\alpha = 0.05$ level of significance in both groups. It should be noted that two different choices on the parameter determining the exponential distribution on group 1 lead to the same result in the likelihood ratio test statistics in group 2 and group 3. In addition, one should be highlighted about the scale parameter λ in the exponential distribution. On

group 1, if the value of the parameter λ determining the exponential distribution is assumed to one, then the parameter estimate is found to be 1.6612 in group 2. The quotient (Q1) of the form $\hat{\lambda}_{gp.2}/\lambda_{gp.1}$ is simply equal to 1.6612. Concerning the another λ=2 imposed on group 1, the parameter estimate is found to be 3.3225 in group 2. The quotient (Q2) of the form $\hat{\lambda}_{gp.2}/\lambda_{gp1}$ is actually equal to 1.6612, just the same as (Q1). This finding should be true since the parameter λ is the scale parameter in the exponential distribution. Similarly, the result also holds in group 3. Therefore, the value of the parameter chosen in the reference distribution is not of importance, it can be arbitrarily taken for comparison convenience.

From the above findings, it is shown that the exponential distribution doesn't fit the data. It is not surprising since the case considered is $s = 3$ which is greater than 2. In § 3.1, we have asserted that the exponential distribution may or may not fit the data if the case of $s > 2$ happens. So, it is required to consider another distribution instead of exponential distribution.

3.5.1.2 Normal distribution

In this section, a normal distribution is imposed on group 1. The value of the parameter which determines the normal distribution is given as $\mu = 0$ and $\sigma = 1$. The thresholds α_2 and α_3 are estimated as :

$$\hat{\alpha_2} = -2.0141, \quad \hat{\alpha_3} = 1.9600.$$

By using the thresholds estimated in group 1, the value of the parameter in the normal distribution on group 2 and group 3 is estimated as :

$$\hat{\mu}_{gp.2} = -0.0312, \quad \hat{\sigma}_{gp.2} = 1.9132,$$

$$\hat{\mu}_{gp.3} = 0.8529, \quad \hat{\sigma}_{gp.3} = 2.0112.$$

To test :

$$H_0 : \text{The data fits the reference distribution} \quad \text{vs.} \quad H_1 : \text{not } H_0.$$

It is required to find the test statistics in group 2 and 3. The likelihood ratio test statistics λ^* are calculated from (3.12). Also, under H_0, it can be shown that,

$$G^2 = -2\log\lambda^* \xrightarrow{l} \chi^2_{d.f.},$$

with $d.f. = s - 3 = 0$. Actually, if we calculate the likelihood ratio test statistics λ^* in group 2 and group 3, both give the value of one. Thus the $G^2 = -2\log\lambda^*$ are equal to zero in both groups.

As an illustration, we are going to see what will happen if the value of the parameter determining the normal distribution on group 1 is given as $\mu = 1$ and $\sigma = 1.5$. By maximum likelihood method, the thresholds α_2 and α_3 are estimated as :

$$\hat{\alpha_2} = -2.0211, \quad \hat{\alpha_3} = 3.9399.$$

By using the thresholds estimated in group 1, the value of the parameter in the normal distribution on group 2 and group 3 is estimated as :

$$\hat{\mu}_{gp.2} = 0.9532, \quad \hat{\sigma}_{gp.2} = 2.8698,$$

$$\hat{\mu}_{gp.3} = 2.2794, \quad \hat{\sigma}_{gp.3} = 3.0167.$$

Then the likelihood ratio test statistics λ^* are found to be one in both groups. From the above results, it is noticed that if the case of $s = 3$ happens, normal distribution can fit the data perfectly.

In addition, one more thing should be noted on the location parameter μ and the scale parameter σ in the normal distribution. On group 1, if the value of the parameter determining the normal distribution is given as $\mu = 0$ and $\sigma = 1$, then the parameter estimate is found to be $\hat{\mu} = -0.0312$, $\hat{\sigma} = 1.9132$ in group 2. Concerning the another $\mu = 1$ and $\sigma = 1.5$ imposed on group 1, the parameter estimate is found to be $\hat{\mu} = 0.9532$, $\hat{\sigma} = 2.8698$ in group 2. Both cases give the same value of -0.0312 under the transformation of $(\hat{\mu} - \mu)/\sigma$. It is due to the fact that μ is the location parameter in the normal distribution. Besides, both also give the same value of quotient with the form $\hat{\sigma}/\sigma$. The quotient is found to be 1.9132. This results from σ is the scale parameter in the normal distribution. Similarly, the above results are also true in group 3.

3.5.1.3 Weibull distribution

Considering the Weibull distribution, if the value of the parameter involved is given by $\gamma = 2, \beta = 6$ on group 1. By maximum likelihood method, the thresholds are estimated as :

$$\hat{\alpha_2} = 0.3653, \quad \hat{\alpha_3} = 4.7046.$$

By fixed at the thresholds estimated above, we then estimate the value of the parameter in Weibull distribution on group 2 and group 3. The result is :

$$\hat{\gamma}_{gp.2} = 0.9630, \quad \hat{\beta}_{gp.2} = 2.3334,$$

$$\hat{\gamma}_{gp.3} = 1.0702, \quad \hat{\beta}_{gp.3} = 4.2486.$$

To test the goodness-of-fit, we also compute the likelihood ratio test statistic λ^* by (3.15). The calculated λ^* are one in both groups and the $G^2 = -2\log\lambda^*$ are equal to zero. It is also the case of perfect fit. Therefore, Weibull distribution can fit the data perfectly.

Next, by considering the value of the parameter determining the Weibull distribution in group 1 as $\gamma = 2$ and $\beta = 4$, the thresholds can be estimated as :

$$\hat{\alpha_2} = 0.2983, \quad \hat{\alpha_3} = 3.8413.$$

By assuming the thresholds are fixed, the value of the parameter in Weibull distribution on group 2 and group 3 are estimated as :

$$\hat{\gamma}_{gp.2} = 0.9630, \quad \hat{\beta}_{gp.2} = 1.9196,$$

$$\hat{\gamma}_{gp.3} = 1.0702, \quad \hat{\beta}_{gp.3} = 3.4199.$$

To test the goodness-of-fit, the likelihood ratio test statistics λ^* are also found to be one in both groups. So, Weibull distribution can fit the data perfectly. With the above two conditions considered, it reveals that if $s = 3$, Weibull distribution can fit the data well. In addition, it is worth considering the shape parameter γ and the scale parameter $\beta^{1/\gamma}$ in the Weibull distribution. On group 1, if the value of the parameter is given as $\gamma = 2$ and $\beta = 6$, the parameter estimate is found to be $\hat{\gamma} = 0.9630$, $\hat{\beta} = 2.3334$ in group 2. Next, if the another $\gamma = 2$, $\beta = 4$ is imposed on group 1, the parameter estimate is found to be $\hat{\gamma} = 0.9630$, $\hat{\beta} = 1.9196$ in group 2. Both cases give the same value of the quotient form $\hat{\beta}^{1/\hat{\gamma}}/\beta^{1/\gamma}$. The calculated value is 0.9842. Similarly, the same result is obtained in group 3. This results from the effect of the scale parameter.

3.5.2 Case 2 ($s = 4$)

Now suppose the number of thresholds is equal to 3. The threshold parameter is $\boldsymbol{\alpha} = \{\alpha_2, \alpha_3, \alpha_4\}$ with known α_1 and α_5. Assume that the true values of thresholds $\alpha_2, \alpha_3, \alpha_4$ are equal to -2, 0, $+2$ respectively. The new table is presented in Table 3.2. Similar to case 1, group 1 is chosen as the reference group. The goodness-of-fit test in three distributions are discussed.

Table 3.2: Artificial 3×4 Contingency Table.

	Category				
	1	2	3	4	Total
Group 1	22	475	478	25	1000
Group 2	150	374	327	149	1000
Group 3	77	264	368	291	1000

3.5.2.1 Exponential distribution

Initially, the value of the parameter λ which determines the exponential distribution is given to one. On group 1, the thresholds $\alpha_2, \alpha_3, \alpha_4$ are estimated by maximum likelihood method. The result is :

$$\hat{\alpha_2} = 0.0222, \quad \hat{\alpha_3} = 0.6872, \quad \hat{\alpha_4} = 3.6889.$$

Assuming the thresholds found are the fixed constants, the value of the parameter in the exponential distribution is estimated on group 2 and group 3 as :

$$\hat{\lambda}_{gp.2} = 1.4469, \quad \hat{\lambda}_{gp.3} = 2.6329.$$

To test whether the data fits the reference distribution, the likelihood ratio test statistic λ^* can be calculated by (3.7). Under the null hypothesis of the test,

H_0 : The data fits the reference distribution,

we have,

$$G^2 = -2\log\lambda^* \xrightarrow{l} \chi^2_{d.f.},$$

with $d.f. = s-2 = 2$. The calculated G^2 is 568.16 in group 2 and 271.05 in group 3. Take $\alpha = 0.05$, both G^2 are much greater than $\chi^2_{2,0.05} = 5.991$. Therefore, H_0 are rejected at $\alpha = 0.05$ level of significance in both groups.

In addition, suppose the value of the parameter determining the exponential distribution on group 1 is taken as two. The estimation of the thresholds in group 1 and the subsequent estimation of the parameter in group 2 and group 3 can also be done by maximum likelihood method. The result is :

$$\hat{\alpha_2} = 0.0445, \quad \hat{\alpha_3} = 1.3743, \quad \hat{\alpha_4} = 7.3778,$$

and

$$\hat{\lambda}_{gp.2} = 2.8938, \qquad \hat{\lambda}_{gp.3} = 5.2658.$$

The calculated likelihood ratio test statistics G^2 are 568.16 in group 2 and 271.05 in group 3. This is a satisfactory result which is the same as the parameter value chosen as one. From those results, it is shown that the exponential distribution doesn't fit the data.

Concerning with the scale parameter λ in the exponential distribution, one can observe that the same values of $\hat{\lambda}_{gp.2}/\lambda$ are obtained with two different choices of λ in group 1. This is also true for group 3's results.

3.5.2.2 Normal distribution

On group 1, the value of the parameter which determines the normal distribution is given as $\mu = 0$ and $\sigma = 1$. The thresholds α_2, α_3 and α_4 are estimated as :

$$\hat{\alpha_2} = -2.0141, \quad \hat{\alpha_3} = -0.0075, \quad \hat{\alpha_4} = 1.9600.$$

By using the thresholds estimated in group 1, the value of the parameter (μ, σ) on group 2 and group 3 is estimated as :

$$\hat{\mu}_{gp.2} = -0.0704, \quad \hat{\sigma}_{gp.2} = 1.9124,$$

$$\hat{\mu}_{gp.3} = 0.8402, \quad \hat{\sigma}_{gp.3} = 2.0189.$$

To test :

$$H_0 : \text{The data fits the reference distribution} \quad \text{vs.} \quad H_1 : \text{not } H_0.$$

The likelihood ratio test statistic λ^* is 0.4262 in group 2 and 0.8917 in group 3. Under H_0,

$$G^2 = -2 \log \lambda^* \xrightarrow{l} {\chi^2}_{d.f.},$$

with $d.f. = s - 3 = 1$. The calculated G^2 is 1.7055 in group 2 and 0.2292 in group 3. Both of them are smaller than ${\chi^2}_{1,0.05} = 3.841$. Therefore, H_0 are not rejected at $\alpha = 0.05$ level of significance in both groups. Similarly, if the value of the parameter which determines the normal distribution in group 1 is given as $\mu = 1$

and $\sigma = 1.5$. The estimation of the thresholds and the subsequent estimation of the parameter are also done by maximum likelihood method. The result is :

$$\hat{\alpha_2} = -2.0211, \quad \hat{\alpha_3} = 0.9887, \quad \hat{\alpha_4} = 3.9399,$$

and

$$\hat{\mu}_{gp.2} = 0.8944, \quad \hat{\sigma}_{gp.2} = 2.8685,$$

$$\hat{\mu}_{gp.3} = 2.2602, \quad \hat{\sigma}_{gp.3} = 3.0284.$$

To test the same hypothesis just mentioned before, It gives the same values in the likelihood ratio test statistics in groups 2 and 3. It tells us that under $s = 4$, normal distribution can fit the data well at certain level of significance.

As a comment, one also observe that the same values of $(\hat{\mu} - \mu)/\sigma$ and $\hat{\sigma}/\sigma$ are obtained in groups 2 and 3 with different choices of μ and σ in group 1. This results from μ and σ are location and scale parameters in the normal distribution.

3.5.2.3 Weibull distribution

In case 1, we have shown that under $s = 3$, Weibull distribution can fit the data perfectly. Now, we consider the case of $s = 4$. On group 1, the value of the parameter (γ, β) is given as $\gamma = 2, \beta = 6$. The thresholds estimated are :

$$\hat{\alpha_2} = 0.3653, \quad \hat{\alpha_3} = 2.0305, \quad \hat{\alpha_4} = 4.7046.$$

By fixed at the thresholds estimated, the value of the parameter is estimated on group 2 and group 3. The result is :

$$\hat{\gamma}_{gp.2} = 0.9940, \quad \hat{\beta}_{gp.2} = 2.5307,$$

$$\hat{\gamma}_{gp.3} = 1.1520, \quad \hat{\beta}_{gp.3} = 4.9288.$$

To do the same things, test :

$$H_0 : \text{The data fits the reference distribution} \quad \text{vs.} \quad H_1 : \text{not } H_0.$$

The likelihood ratio test statistic λ^* is 0.0051 in group 2 and 0.0012 in group 3. Under H_0,

$$G^2 = -2\log\lambda^* \xrightarrow{l} \chi^2_{d.f.},$$

with $d.f. = s - 3 = 1$. The calculated G^2 is 10.5601 in group 2 and 13.4150 in group 3. Both of them are greater than $\chi^2_{1,0.05} = 3.841$. Therefore, H_0 are rejected at $\alpha = 0.05$ level of significance. Similarly, if the value of the parameter which determines the Weibull distribution in group 1 is given as $\gamma = 2$ and $\beta = 4$. By maximum likelihood method, the estimation of the thresholds and the subsequent estimation of the parameter are done. The result is :

$$\hat{\alpha_2} = 0.2983, \quad \hat{\alpha_3} = 1.6579, \quad \hat{\alpha_4} = 3.8413,$$

and

$$\hat{\gamma}_{gp.2} = 0.9940, \quad \hat{\beta}_{gp.2} = 2.0688,$$

$$\hat{\gamma}_{gp.3} = 1.1520, \quad \hat{\beta}_{gp.3} = 3.9023.$$

Finally, it also gives an identical result in the likelihood ratio test statistics. Concerning with the scale parameter $\beta^{1/\gamma}$ in the Weibull distribution, the same values of $\hat{\beta}^{1/\hat{\gamma}}/\beta^{1/\gamma}$ are obtained in groups 2 and 3 with different choices of β in group 1. This is similar in § 3.5.1.3.

Combined with the results obtained in case 2 ($s = 4$), exponential distribution and Weibull distribution can't fit the data. Only normal distribution does fit at certain level of significance ! In case 1 ($s = 3$), exponential distribution can't fit the data while normal distribution and Weibull distribution fit the data perfectly. It agrees with the assertions given by the last paragraphs in sections 3.1, 3.2 and 3.3. In general, it is not restricted in those three distributions. Many other distributions may be possible to be considered. Therefore, it is worth noting the generalization of the result on the goodness-of-fit test discussed in § 3.4.

Chapter 4

Real Data Illustration

In this thesis, the objective is to illustrate the test of homogeneity in distributions with ordinal categories. If the case of non-homogeneity happens, it is possible to examine the heterogeneity among the distributions in a relative sense. Returning to the data given in Table 2.1, subjects were asked whether courts were sufficiently harsh with criminals. From the year 1972 to 1974, four independent samples were obtained. To examine whether there is a trend over the years of survey, the test of homogeneity of four independent samples is worth to be investigated. Basically, the response is classified with five categories, that is, "Too harshly", "About right", "Not harshly enough", "Don't know" and "No answer". Since the ordinal scales are considered only, the interesting categories are "Not harshly enough", "About right" and "Too harshly". The renewed table is presented in Table 4.1.

Table 4.1: Modified Data Set with Ordered Responses.

Year	Response			
of survey	Not harshly enough	About right	Too harshly	Total
1972	1066	265	105	1436
1973	1092	196	68	1356
1974	580	72	42	694
1975	1174	144	61	1379
Total	3912	677	276	4865

In general, we first consider an $I \times J$ contingency table with I and J being the number of rows and columns respectively. We use the symbols A and B for the two variables. Let n_{ij} be the cell frequency, $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$ with $\sum_i \sum_j n_{ij} = n$ being the total sample size. Also, let $\pi_{j(i)}$ be the conditional probability of level j of B at level i of A. If a sampling scheme that fixes row totals is used, then the general hypothesis for testing of homogeneity (in row distributions across columns) is :

$$H_0 : \pi_{j(1)} = \pi_{j(2)} = \cdots = \pi_{j(I)}, \qquad j = 1, 2, \ldots, J,$$

vs. H_1: not H_0.

In tradition, the Pearson chi-squared statistic is commonly used to test homo-

geneity. When the homogeneity holds, the Pearson chi-squared statistic,

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i+}n_{+j}}{n})^2}{\frac{n_{i+}n_{+j}}{n}},$$

has an asymptotic chi-squared distribution with degrees of freedom, $d.f. = (I - 1)(J - 1)$.

Referring to Table 4.1, let us examine the data by classical approach firstly. Based on the reference year 1972, the parallel tests of homogeneity are constructed as :

$$H_0 : \pi_{j(1)} = \pi_{j(2)}, \quad j = 1,2,3 \qquad \text{vs. } H_1 : \text{not } H_0,$$

$$H_0 : \pi_{j(1)} = \pi_{j(3)}, \quad j = 1,2,3 \qquad \text{vs. } H_1 : \text{not } H_0,$$

$$\text{and } H_0 : \pi_{j(1)} = \pi_{j(4)}, \quad j = 1,2,3 \qquad \text{vs. } H_1 : \text{not } H_0.$$

Under the null hypothesis of homogeneity, the chi-squared statistic has an asymptotic chi-squared distribution with degrees of freedom, $d.f. = (2-1)(3-1) = 2$ in each test. The calculated chi-squared statistics are 16.2752, 25.6614 and 51.5338 in sequence. Compared with the critical value, $\chi^2_{2,0.05} = 5.991$, it concludes that three null hypotheses of homogeneity are rejected at 0.05 level of significance. Thus we come to an end that the distribution of attitudes toward the courts appears to have changed in the period of time under study. In this situation, it is intuitively possible to examine the heterogeneity among the distributions in a relative sense by the current method proposed in this thesis.

In Table 4.1, the number of categories is only three in the response variable.

They are ordered in the sequence: "Not harshly enough", "About right" and "Too harshly". Therefore, the thresholds considered are α_2, α_3 with known α_1 and α_4. The number of thresholds is two in this case. From the four years of survey, it is necessary to impose a univariate distribution on the reference year and then study the distribution on each remaining year relatively. Without loss of generality, we take 1972 as the reference year. Based on the results obtained in Chapter 3, the most suitable family of distribution should have only two independent parameters. It leads to the case of perfect fit of the data set by this kind of distribution. However, as an illustration, the goodness-of-fit tests are carried out on all three distributions currently discussed in this thesis, that is, exponential distribution, normal distribution and Weibull distribution.

4.1 Test for the Exponential distribution

In this section, an exponential distribution is imposed on the reference year, 1972. The value of the parameter λ which determines the exponential distribution is assumed to one. By maximum likelihood method, the thresholds α_2 and α_3 are estimated as :

$$\hat{\alpha_2} = 1.3561, \quad \hat{\alpha_3} = 2.6157.$$

By assuming the thresholds estimated are the fixed constants, the value of the parameter in the exponential distribution can also be estimated on the remaining years, 1973, 1974 and 1975. The result is :

$$\hat{\lambda}_{1973} = 0.8446, \quad \hat{\lambda}_{1974} = 0.8134, \quad \hat{\lambda}_{1975} = 0.7510.$$

Consider testing :

$$H_0 : \text{ The data fits the reference distribution} \quad \text{vs.} \quad H_1 : \text{ not } H_0.$$

The likelihood ratio test statistic, λ^* can be calculated by (3.7). Under H_0,

$$G^2 = -2\log\lambda^* \xrightarrow{l} \chi^2_{d.f.},$$

with $d.f. = 1$. The calculated G^2 are 1.8617, 17.2562 and 17.1401 in the years 1973, 1974 and 1975 respectively. Compared with the critical value, $\chi^2_{1,0.05} = 3.841$, we notice that only the G^2 found in the year 1973 is smaller than 3.841 while the other two are larger than 3.841. It indicates that the exponential distribution is only suitable to handle the data set in years 1972 and 1973.

4.2 Test for the Normal distribution

On the reference year 1972, suppose a normal distribution with $\mu = 0, \sigma = 1$ is imposed. By maximum likelihood method, the thresholds are estimated as :

$$\hat{\alpha_2} = 0.6506, \quad \hat{\alpha_3} = 1.4529.$$

By using the thresholds estimated in the reference year, the value of the parameter in the normal distribution is estimated in the remaining years as :

$$\hat{\mu}_{1973} = -0.2318, \quad \hat{\sigma}_{1973} = 1.0251,$$

$$\hat{\mu}_{1974} = -0.7167, \quad \hat{\sigma}_{1974} = 1.3994,$$

$$\hat{\mu}_{1975} = -0.6139, \quad \hat{\sigma}_{1975} = 1.2133.$$

Consider testing :

$$H_0 : \text{ The data fits the reference distribution} \quad \text{vs.} \quad H_1 : \text{ not } H_0.$$

The likelihood ratio test statistic, λ^* is calculated from (3.12). Not surprisingly, the calculated λ^* are all equal to one in years 1973, 1974 and 1975. It is the perfect fit case since there are two independent parameters involved in normal distribution. Therefore, normal distribution is the suitable one to handle the whole data set given in Table 4.1.

4.3 Test for the Weibull distribution

The last distribution to be considered is the Weibull distribution. Firstly, we impose a Weibull distribution with $\gamma = 2, \beta = 6$ on the reference year. To do the same thing, the thresholds are estimated as :

$$\hat{\alpha_2} = 2.8525, \quad \hat{\alpha_3} = 3.9616.$$

By using the thresholds estimated in the reference year, the value of the parameter in the Weibull distribution is estimated in the remaining years as :

$$\hat{\gamma}_{1973} = 1.8382, \quad \hat{\beta}_{1973} = 4.1966,$$

$$\hat{\gamma}_{1974} = 1.3399, \quad \hat{\beta}_{1974} = 2.2550,$$

$$\hat{\gamma}_{1975} = 1.4986, \quad \hat{\beta}_{1975} = 2.5237.$$

The likelihood ratio test statistic, λ^* given by (3.15) can be evaluated in testing whether the Weibull distribution fits the data set. The calculated λ^* are all equal to one in the three years. Similar to normal distribution, it is the case of perfect fit. Therefore, Weibull distribution can also fit the data set well.

As a clarification, since the number of the thresholds considered in this case is two, which is the same as the number of independent parameters involved in the normal distribution or Weibull distribution, so the data set presented in Table 4.1 can be perfectly fitted by both distributions.

4.4 Inferences from the Exponential distribution

In § 4.1, it is found that the exponential distribution is suitable to study the difference between the reference year 1972 and year 1973. On the reference year, an exponential distribution with $\lambda = 1$ is arbitrarily imposed to it. By maximum likelihood estimation, the thresholds are estimated. By fixing the thresholds estimated, the value of the parameter in the exponential distribution on year 1973 is estimated as $\hat{\lambda}_{1973} = 0.8446$ with the estimated standard error, $\hat{\text{S.E.}}(\hat{\lambda}_{1973}) = 0.0261$. With reference in year 1972, we set the hypothesis as :

$$H_0 : \lambda = 1 \qquad \text{vs.} \quad H_1 : \lambda \neq 1,$$

to compare the difference. The results are presented in Table 4.2.

Table 4.2: Hypothesis Testing on parameter λ.

	Hypothesis	χ^2-statistic	$d.f.$	P-value
1973	$H_0 : \lambda = 1$ vs. $H_1 : \lambda \neq 1$	35.4995	1	0.0000

From the above results, it is observed that the p-value is smaller than 0.05 or even 0.01. Hence, the case of non-homogeneity between the years 1972 and 1973 happens. In the following sections, we will analyze the data set by normal and Weibull distributions.

4.5 Inferences from the Normal distribution

Just before, it has found that the data set in Table 4.1 is perfectly fitted by the normal distribution. We are going to set up hypothesis on the parameters estimated to study the location and variation differences on the years of survey relatively.

As in § 2.2.2, let us denote $\boldsymbol{\theta} = (\mu, \sigma)'$ containing the parameters in the normal distribution. If $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma})'$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, it is known that,

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{l} N[0,\ \mathrm{I}^{-1}(\boldsymbol{\theta})],$$

holds according to the asymptotic theory. Since Hessian matrix converges in probability to $\mathrm{I}(\boldsymbol{\theta})$, then the Hessian matrix can be used to approximate the information matrix. The maximum likelihood estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by $(H(\hat{\boldsymbol{\theta}}))^{-1}$. Therefore, the estimated standard error of the estimated parameter can be obtained.

In § 4.2, we have imposed a normal distribution with $\mu = 0, \sigma = 1$ on the reference year, 1972. By fixing the thresholds estimated, the value of the parameter in the normal distribution is then estimated in the remaining years. Together with the estimated standard errors of the parameter estimates, the results are presented in Table 4.3. With reference in the year 1972, it is necessary

Table 4.3: Parameter Estimates with Estimated Standard Errors (S.E.).

	Parameter	Parameter Estimate	Estimated S.E.
1973	μ	-0.2318	0.0629
	σ	1.0251	0.0547
1974	μ	-0.7168	0.1482
	σ	1.3994	0.1237
1975	μ	-0.6139	0.0905
	σ	1.2133	0.0717

to test hypotheses as :

$$H_0 : (\mu, \sigma^2)' = (0,1)' \quad \text{vs. } H_1 : \text{not } H_0,$$

$$H_0 : \mu = 0 \qquad \text{vs. } H_1 : \mu \neq 0,$$

and

$$H_0 : \sigma^2 = 1 \qquad \text{vs. } H_1 : \sigma^2 \neq 1,$$

on the remaining years to compare their means and variances. The results are presented in Table 4.4.

From the results on the hypothesis testing, it is directly observed that the p-values are all smaller than 0.05 or even 0.01 except the one on testing H_0 : $\sigma^2 = 1$ in year 1973. Hence, the sources of heterogeneity are due to location

Table 4.4: Hypothesis Testing on parameters, μ and σ.

	Hypothesis	χ^2-statistic	$d.f.$	P-value
1973	$H_0 : (\mu, \sigma^2)' = (0,1)'$ vs. H_1 : not H_0	29.0318	2	0.0000
	$H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$	13.5912	1	0.0002
	$H_0 : \sigma^2 = 1$ vs. $H_1 : \sigma^2 \neq 1$	0.2165	1	0.6417
1974	$H_0 : (\mu, \sigma^2)' = (0,1)'$ vs. H_1 : not H_0	23.6159	2	0.0000
	$H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$	23.3955	1	0.0000
	$H_0 : \sigma^2 = 1$ vs. $H_1 : \sigma^2 \neq 1$	15.0031	1	0.0001
1975	$H_0 : (\mu, \sigma^2)' = (0,1)'$ vs. H_1 : not H_0	65.7932	2	0.0000
	$H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$	46.0178	1	0.0000
	$H_0 : \sigma^2 = 1$ vs. $H_1 : \sigma^2 \neq 1$	10.8305	1	0.0010

difference and dispersion difference on the year 1974 and the year 1975 with reference in year 1972. Compared with the year 1973, the non-homogeneity is due to location difference rather than dispersion difference. This finding can provide some insights in analyzing the testing of homogeneity in categorical data with ordinal variables.

4.6 Inferences from the Weibull distribution

Besides considering the normal distribution, the data set can also be analyzed by the Weibull distribution. In § 4.3, a Weibull distribution with $\gamma = 2, \beta = 6$ is arbitrarily placed on the reference year, 1972. The estimation of the thresholds on this year's data and the subsequent estimation of the parameter in the Weibull distribution on the following years are done by maximum likelihood method. Similar to the previous section, the parameter estimates with their estimated standard errors are presented in Table 4.5.

Table 4.5: Parameter Estimates with Estimated Standard Errors (S.E.).

	Parameter	Parameter Estimate	Estimated S.E.
1973	γ	1.8382	0.1161
	β	4.1966	0.5820
1974	γ	1.3398	0.1427
	β	2.2550	0.3888
1975	γ	1.4986	0.1122
	β	2.5238	0.3367

To study the differences among the years of survey, it is intuitively to test

the hypotheses as :

$$H_0 : \ (\gamma, \beta)' = (2, 6)' \qquad \text{vs. } H_1 : \text{ not } H_0,$$

$$H_0 : \ \gamma = 2 \qquad \text{vs. } H_1 : \ \gamma \neq 2,$$

and

$$H_0 : \ \beta = 6 \qquad \text{vs. } H_1 : \ \beta \neq 6.$$

The results are presented in Table 4.6.

Table 4.6: Hypothesis Testing on parameters, γ and β.

	Hypothesis	χ^2-statistic	d.f.	P-value
1973	$H_0 : (\gamma, \beta)' = (2, 6)'$ vs. H_1: not H_0	62.8134	2	0.0000
	$H_0 : \gamma = 2$ vs. $H_1 : \gamma \neq 2$	1.9449	1	0.1631
	$H_0 : \beta = 6$ vs. $H_1 : \beta \neq 6$	9.6030	1	0.0019
1974	$H_0 : (\gamma, \beta)' = (2, 6)'$ vs. H_1: not H_0	426.2483	2	0.0000
	$H_0 : \gamma = 2$ vs. $H_1 : \gamma \neq 2$	21.4050	1	0.0000
	$H_0 : \beta = 6$ vs. $H_1 : \beta \neq 6$	92.7947	1	0.0000
1975	$H_0 : (\gamma, \beta)' = (2, 6)'$ vs. H_1: not H_0	658.1984	2	0.0000
	$H_0 : \gamma = 2$ vs. $H_1 : \gamma \neq 2$	19.9855	1	0.0000
	$H_0 : \beta = 6$ vs. $H_1 : \beta \neq 6$	106.5852	1	0.0000

The results in Table 4.6 may somehow similar to those in Table 4.4. All the

p-values found are very small except the one on testing $H_0 : \gamma = 2$ in year 1973. This may indicate the year 1972 and year 1973 have the same shape. By using Weibull distribution to analyze the data set, it is also helpful to contribute to the source of heterogeneity across several independent samples. In other words, given a data set, once a suitable distribution is in hand, we can study the source of heterogeneity in a relative sense.

Chapter 5

Conclusion

In this thesis, our objective is to test homogeneity in distributions with ordinal categories. Within several independent samples, it is required to take one as the reference group without loss of generality. Firstly, a univariate distribution with known parameter value is arbitrarily imposed on the reference group. This univariate distribution is suitably chosen depending on the number of the thresholds and the number of the parameters involved. The details have been discussed in Chapter 3. Based on the reference group, the thresholds can be estimated by maximum likelihood method. By assuming those threshold estimates are the fixed constants, the value of the parameter in the same distribution of each remaining group can also be estimated by maximum likelihood method. Therefore, it is possible to compare the distribution of each remaining group with the refer-

ence group relatively by setting up useful hypotheses on the parameters. By using this proposed method, it is helpful to study the homogeneity in distributions with ordinal categories.

Although only three distributions are considered in this thesis, it is also possible to consider other common distributions such as lognormal distribution, double exponential distribution and so on. In real life, it may not always be the case of independent samples. It is possible to extend the above idea to the case of dependent samples. Common example is bivariate normal distribution or even consider bivariate elliptical distribution.

Bibliography

[1] Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* New York: John Wiley.

[2] Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley.

[3] Balakrishnan, N., Johnson, N.L. and Kotz, S. (1994). *Continuous Univariate Distributions: Vol. 1.* New York: John Wiley.

[4] Bard, Y. (1974). *Nonlinear Parameter Estimates.* New York: Academic.

[5] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press.

[6] Eliason, S.R. (1993). *Maximum Likelihood Estimation: Logic and Practice.* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 096). Newbury Park, CA: Sage.

[7] General Social Surveys, 1972-1982: *Cumulative Codebook.* National Opinion Research Center: Chicago.

[8] Haberman, S.J. (1978). *Analysis of Qualitative Data: Vol. 1. Introductory Topics.* New York: Academic Press.

[9] IMSL Library. (1991). *International Mathematical and Statistical Libraries. (Ver 2.0).* Houston, Texas.

[10] Lee, S.Y. and Poon, W.Y. (1986). Maximum likelihood estimation of polyserial correlations. *Psychometrika, 51*, 113-121.

[11] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* New York: Wiley.